SECURITY CLASSIFICATION OF THIS PAGE

DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

AD-A218 359

DTIC ELECTE FEB 26 1990 D B

1b. RESTRICTIVE MARKINGS	
2b. DECLASSIFICATION / DOWNGRADING SCHEDULE	3. DISTRIBUTION / AVAILABILITY OF REPORT APPROVED FOR PUBLIC RELEASE: DISTRIBUTION UNLIMITED
4. PERFORMING ORGANIZATION REPORT NUMBER(S)	5. MONITORING ORGANIZATION REPORT NUMBER(S) AFOSR.TR. 90-0231
6a. NAME OF PERFORMING ORGANIZATION DEPARTMENT OF PSYCHOLOGY STANFORD UNIVERSITY 6b. OFFICE SYMBOL (If applicable)	7a. NAME OF MONITORING ORGANIZATION AIR FORCE OFFICE OF SCIENTIFIC RESEARCH (NL)
6c. ADDRESS (City, State, and ZIP Code) BUILDING 420, JORDAN HALL STANFORD, CA 94305-2130	7b. ADDRESS (City, State, and ZIP Code) BOLLING AIR FORCE BASE, DC 20332-6448
8a. NAME OF FUNDING / SPONSORING ORGANIZATION AFOSR 8b. OFFICE SYMBOL (If applicable) NL	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER AFOSR-89-0064
8c. ADDRESS (City, State, and ZIP Code) BUILDING 410 BOLLING AFB, DC 20332-6448	10. SOURCE OF FUNDING NUMBERS PROGRAM ELEMENT NO. 61102F; PROJECT NO. 2313; TASK NO. A4; WORK UNIT ACCESSION NO.

11. TITLE (Include Security Classification)
DECISION UNDER CONFLICT: RESOLUTION AND CONFIDENCE IN JUDGMENT AND CHOICE

12. PERSONAL AUTHOR(S)
TVERSKY, AMOS

13a. TYPE OF REPORT	13b. TIME COVERED	14. DATE OF REPORT (Year, Month, Day)	15. PAGE COUNT
FIRST ANNUAL TECHNICAL	FROM 11-1-88 TO 11-30-89	JANUARY 18, 1990	18

16. SUPPLEMENTARY NOTATION

17. COSATI CODES		
FIELD	GROUP	SUB-GROUP
05	09	

18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
compatibility, evidence, ambiguity, competence.

19. ABSTRACT (Continue on reverse if necessary and identify by block number)

A constructive approach to the analysis of judgment and choice maintains that the decision maker does not always have well-defined preferences and beliefs. Instead, they are often constructed in the elicitation process. This approach is used to explain and interpret a variety of phenomena that violate the classical theory of rational choice. It also leads to the formulation of psychological principles that govern judgment and choice.

The present report summarizes three research projects conducted within the constructivist framework. The first project investigates the compatibility principle according to which the weighting of a stimulus attribute is enhanced by its compatibility with the response. This principle is applied to the analysis of the well-known preference reversal phenomenon, namely the observation that the preference order between bets does not agree with

(over)

20 DISTRIBUTION / AVAILABILITY OF ABSTRACT	21. ABSTRACT SECURITY CLASSIFICATION
☒ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT. ☐ DTIC USERS	UNCLASSIFIED
22a. NAME OF RESPONSIBLE INDIVIDUAL ALFRED R. FREGLY, PH.D.	22b. TELEPHONE (Include Area Code) (202) 767-5021
	22c. OFFICE SYMBOL NL

DD Form 1473, JUN 86 — Previous editions are obsolete. — SECURITY CLASSIFICATION OF THIS PAGE

19. ABSTRACT

the ordering of their selling prices. Preference reversals are explained by the fact that since both the payoffs and the prices are expressed in dollars, the payoffs are weighted more heavily in pricing than in choice. The second project investigates the determinants of confidence in the evaluation of evidence. We argue that the pattern of overconfidence and underconfidence observed in the literature is produced by the relations of dominance between components of evidence. In particular, overconfidence arises when the strength of evidence is high but its weight is low, and underconfidence occurs when strength is low and weight is high. The third project concerns the relation between judgments of belief and preference between bets. We show that people's willingness to bet on their beliefs depends not only on the perceived likelihood of the event in question (subjective probability) and the precision with which it is measured (ambiguity or vagueness); it also depends on people's general knowledge or competence regarding the relevant domain. We hypothesize that, holding belief or judged probability constant, people prefer to bet in areas where they feel knowledgeable or competent and they avoid betting in domains where they feel ignorant or uninformed. This account, called the competence hypothesis, is supported in a series of experiments showing that people prefer to bet on their belief over a matched chance event in their area of expertise and that they prefer to bet on a chance event in areas where they regard themselves as less knowledgeable. This result indicates that the relation between preference and belief is considerably more complicated than implied by traditional models, and that it is exceedingly difficult, if not impossible, to infer beliefs from preferences.

Decision under Conflict: Resolution and Confidence in Judgment and Choice

Annual Technical Report for 1989

The major theme of the present grant is the constructive nature of judgment and choice. According to this view, people do not always have well-defined values, hence preferences and beliefs are often constructed -- not merely revealed -- in the elicitation process. Furthermore, choice is contingent in the sense that different constructions (e.g., descriptions of the options or methods of elicitation) could lead to different responses, contrary to the principle of invariance that underlies the normative theory of judgment and choice. Thus, we propose that many phenomena of choice under both risk and certainty are better understood as an attempt to resolve conflict and reduce cognitive strain rather than as an attempt to maximize a well-defined target function.

During the first year of work, my collaborators and I have focused on three major projects within a general constructive framework: (i) elicitation effects and the compatibility principle, (ii) the weighing of evidence and the determinants of confidence, and (iii) ambiguity and competence in choice under uncertainty. The three projects will be described in turn; further details can be found in the enclosed articles.

Elicitation Effects and the Compatibility Principle

Perhaps the major empirical observation that calls for a constructive approach is the dependence of choice on the method of elicitation. There is a large body of evidence showing that strategically equivalent methods of elicitation give rise to systematically different responses

(see e.g., Slovic, Fischhoff, & Lichtenstein, 1982; Tversky, Sattath, & Slovic, 1988). These data are at variance with the classical theory in which the decision maker has a well-defined preference order that can be elicited using different procedures. To account for these data within a constructive framework, we seek explanatory principles that relate the characteristics of the task to the attributes of the object under study. One such notion is the compatibility hypothesis, which states that the weight of a stimulus attribute is enhanced by its compatibility with the response.

The rationale for this hypothesis is two-fold. First, if the input and the output are non-compatible, additional mental operations are usually required which increase effort and error, and reduce impact. Second, a response mode tends to focus attention on the compatible features of the stimulus. The significance of the compatibility between the input and the output has long been recognized by students of human performance. Engineering psychologists have discovered that responses to visual displays of information, such as an instrument panel, are faster and more accurate if the response structure is compatible with the arrangement of the stimuli. As in the study of perceptual-motor performance, we do not have a formal definition of compatibility or an independent measurement procedures. Nevertheless, in many contexts the compatibility order is sufficiently clear so that it can be investigated experimentally. For example, it seems reasonable to suppose that a turn signal in which a left movement indicates a left turn and a right movement indicates a right turn is more compatible than the opposite design. By comparing people's performance with the two turn signals, it is possible to test whether the more compatible design yields better performance. Similarly, it seems reasonable to assume that the monetary payoffs of a bet are more compatible with a pricing response than with a choice because both

payoffs and prices are expressed in dollars. By comparing choice and pricing, therefore, we can test the hypothesis that the payoffs of a bet loom larger in pricing than in choice.

A simple study conducted in collaboration with Paul Slovic and Dale Griffin (Slovic, Griffin, & Tversky, 1990) illustrates the compatibility principle. Participants predicted the 1987 market value of 12 companies (taken from *Business Week's* top 100) on the basis of their 1986 market value (in billions of dollars), and their rank (among the top 100) with respect to 1987 profits. Half the subjects predicted the 1987 market value in billions of dollars, whereas the other half predicted the companies rank with respect to its 1987 market value. As implied by compatibility, each predictor was given more weight when the criterion was expressed on the same scale (e.g., money, rank). As a consequence, the relative weight of the 1986 market value was twice as high for those who predicted in dollars than for those who predicted the corresponding rank. This effect produced many reversals in which one company was ranked above another but the order of their predicted values was reversed.

Further evidence for the compatibility principle was observed in the prediction of academic performance. The subjects predicted the performance of ten target students in a history course on the basis of the students' performance in two other courses: english literature and philosophy. For each of the ten students, the subjects were given a letter grade (from A+ to D) in one course, and a class rank (from 1 to 100) in the other course. One-half of the subjects predicted the students' grade in history whereas the other half predicted the students' class rank in history. As implied by compatibility, the input variable (english or philosophy) was weighted more heavily when it was expressed in the same unit as the criterion (rank or grade).

The significance of the compatibility principle stems from its ability to explain the well-known preference reversal phenomenon, which has puzzled psychologists and economists for nearly two decades. Subjects are asked to choose between two gambles which have nearly the same expected values. One gamble, called the H bet, has a high chance of winning a relatively small prize (e.g., 8/9 chance to win $4) whereas the other gamble, called the L bet, offers a lower chance to win a larger prize (e.g., 1/9 chance to win $40). Most subjects choose the H bet. Subjects are then asked to state the lowest price at which they would be willing to sell each of the gambles. Surprisingly, people set a higher price for the L bet than for the H bet. In a recent study that used the above pair of bets, for example, 71% of the subjects chose the H bet, but 67% priced L above H. This pattern of preferences, first demonstrated by Lichtenstein and Slovic (1971), has been replicated in numerous studies using monetary incentives and several procedural variations.

Although the preference reversal (PR) phenomenon has been established in many studies, it causes have not been uncovered heretofore. Using a novel experimental design and a new diagnostic procedure, we were able to eliminate the most common explanation of PR as a violation of transitivity, of the independence axiom or of the reduction axiom of expected utility theory. Instead, the great majority of PR patterns are produced by a choice-pricing discrepancy, and more specifically by an overpricing of the L bet (for details see Tversky, Slovic, & Kahneman, 1990). Because the cash equivalence of a bet is expressed in dollars, compatibility implies that the payoffs, which are expressed in the same units, will be weighted more heavily in pricing than in choice. Furthermore, because the payoffs of L bets are much larger than the payoffs of H bets, the major consequence of a compatibility bias is the overpricing of the L bet, which is the

major source of PR.

This account has been supported by several additional experiments. Slovic, Griffin, and Tversky (1990) presented subjects with H and L bets which have non-monetary outcomes, such as a one-week pass for all movie theatres in town or a dinner for two at a good restaurant. If PRs are due primarily to the compatibility of prices and payoffs, which are both expressed in dollars, their incidence should be substantially reduced by the use of non-monetary outcomes. Indeed, the prevalence of PR was reduced by nearly one-half. Further support for the role of compatibility in PR was obtained by Schkade and Johnson (1989), who used a computer-controlled experiment in which the subject can see only one component of each bet at a time. These investigators found that the percentage of time spent looking at the payoff was significantly greater in pricing than in choice. This observation supports the hypothesis that subjects focus their attention on the stimulus components that are most compatible with the response mode.

The compatibility principle also predicts a new type of reversal that has not been investigated heretofore. According to the compatibility hypothesis, the presence of risk is not essential for PR. A discrepancy between choice and pricing should also occur in riskless options with a monetary component, such as delayed payments. Consider the choice between a long-term prospect L (e.g., receiving $2500, 5 years from now) and short-term prospect S (e.g., receiving $1600, 1½ years from now). Suppose now that subjects are asked to choose between L and S, as well as to price these prospects by stating the smallest *immediate* cash payment for which they would be willing to exchange the delayed payment. As in the analysis of the risky case, compatibility implies that the monetary component would loom larger in pricing than in choice. As a consequence, subjects are expected to produce preference reversals in which the short-term

prospect S is preferred over the long-term prospect L in a direct choice, but L is priced higher than S. This prediction was confirmed by Tversky, Slovic, and Kahneman (1990) who presented subjects with pairs of S and L prospects, with comparable present values. The subjects chose between the prospects and also priced each prospect separately. Subjects exhibited the predicted pattern of preference, and the incidence of reversals was no smaller than in the risky case. Overall, the subjects chose the short-term option 74% of the time, but priced the long-term option above the short-term option 75% of the time. Further analysis revealed that, as in the risky case, the major source of reversals was the overpricing of the long-term prospect L, as entailed by compatibility.

Taken together, the experimental and theoretical analyses summarized above demonstrate that scale compatibility plays an important role in both risky and riskless contexts, and that the compatibility principle can explain a wide range of elicitation effects reported in the literature by relating the characteristics of the task to the attribute of the objects under study. For further analysis and discussion, see Slovic, Griffin and Tversky (1990), Tversky, Sattath and Slovic (1988), Tversky, Slovic and Kahneman (1990), and Tversky and Thaler (1990).

The Weighing of Evidence and the Determinants of Confidence

In collaboration with a former graduate student (Dale Griffin), we have completed a series of studies concerning the determinants of confidence. Confidence is important because it controls action. People generally act on beliefs that are held with a high degree of confidence and are reluctant to act in the presence of doubt. Either overconfidence or underconfidence, therefore, can lead to inappropriate action. We have proposed that the pattern of overconfidence

and underconfidence observed in the literature is produced by an improper weighting of evidential components. Judgment under uncertainty typically requires the integration of different items of evidence. The study of human judgment has shown that intuitive weighing of evidence departs systematically from the laws of probability and statistics (see e.g., Kahneman, Slovic, & Tversky, 1982). In particular, some variables such as the strength of evidence (sample proportion, or the warmth of a letter of recommendation) tend to dominate other variables such as the weight of evidence (sample size, or the credibility of the writer). We say that variable A dominates variable B if A is consistently overweighted relative to B, in comparison to the appropriate normative theory.

We have proposed that the strength or extremeness of evidence generally dominates its weight or credence. For example, the subjective probability that a particular experimental result will be replicated is highly sensitive to the size of the effect (e.g., the difference between the means), and not sufficiently sensitive to the sample size, which reflects its weight or credence (Tversky & Kahneman, 1971). If people are highly sensitive to the strength of evidence and not sufficiently sensitive to its weight, then their judgments will be overconfident when strength is high and weight is low, and their judgments will be underconfident when weight is high and strength is low. This account is supported in a series of experimental studies using both random sampling and evidential problems. For example, people overestimated the posterior probability of a binomial hypothesis when sample proportion (strength) was high and sample size (weight) was low, and they underestimated the posterior probability when sample size was large and sample proportion was not extreme. The present results and analysis can help reconcile the apparent disagreement between Edwards' (1968) observation of conservatism in a sequential updating

paradigm, and Tversky and Kahneman's (1971) finding of radical inference in investigators' confidence regarding the replicability of their results. The present account suggests that dominance of sample proportion over sample size is responsible for both findings. In the updating experiments conducted by Edwards and his collaborators, subjects are exposed to large samples of data, typically of moderate strength. This is a context in which underconfidence or conservatism is expected. The situations studied by Tversky and Kahneman, on the other hand, involve moderately strong effects based on fairly small samples. This is a context in which overconfidence is likely to prevail. Both conservative and radical inferences, we propose, are generated by a common bias in the weighing of evidence, namely the dominance of strength over weight.

In many evidential problems, encountered in real life, the evidence cannot be readily decomposed into strength and weight components. Nevertheless, we argue that the strength-weight distinction could be invoked to explain the presence of overconfidence and underconfidence in these problems as well. Note that the strength-weight distinction is closely related to the strategy of prediction by evaluation, investigated by Kahneman and Tversky (1973). Prediction by evaluation leads to overconfidence when people form an extreme impression in a situation where much is unknown (e.g., when an extreme proportion is observed in a small sample). Prediction by evaluation leads to underconfidence when people form a moderate impression on the basis of an extensive body of evidence (e.g., a 5% advantage observed in a large sample). Thus, we suggest that the overconfidence observed in the prediction of future events and the underconfidence observed in the prediction of one's future choice can be explained by the fact that the weight of evidence is greater in the latter than in the former case.

This account has been supported in a study in which subjects were asked to predict their own behavior as well as that of another person, whom they met briefly, in a series of independent prisoner's dilemma games. People exhibited significant overconfidence in predicting the behavior of others, whom they did not know very well, and they were slightly underconfident in predicting their own behavior (Tversky & Griffin, 1990). The presence of overconfidence and underconfidence is important not only because it demonstrates the discrepancy between intuitive judgments and the laws of chance, but primarily because confidence controls action. It has been argued (see e.g., Taylor & Brown, 1988) that overconfidence is adaptive because it moves people to do things they wouldn't have done otherwise. The advantages of overconfidence, however, may be purchased at a high price. Overconfidence in the diagnosis of a patient, the outcome of a trial, or the projected interest rate could lead to inappropriate medical treatment, bad legal advise, and regrettable financial investments. It can be argued that people's willingness to engage in military, legal and other battles would be reduced if they had a more realistic assessment of their chances of success. We doubt that the benefits of overconfidence outweigh its costs.

Ambiguity and Competence in Choice under Uncertainty

The decision we make, conclusions we reach, and the explanations we offer are generally based on beliefs regarding the likelihood of uncertain events such as the outcome of an election, the guilt of a defendant, or the result of a medical operation. In the absence of an objective method for computing the probabilities of such events, we must rely on intuitive judgment as the major instrument for the measurement of uncertainty. There are two general methods for

measuring belief or subjective probability. Perhaps the simplest procedure is to ask the respondent to express his or her belief on a scale from 0 to 100. This task requires a mapping of an impression or a mental state into the language of chance. When we say that the chance of some uncertain event (e.g., rain over the weekend) is 30%, for example, we express the belief that we consider this event to be as probable as a drawing of a red ball from a box that contains 30 red and 70 green balls. Judgments of subjective probability are the most common procedure for measuring belief. However, skepticism about the meaningfulness of verbal responses and the validity of introspection has led decision theorists to devise an alternative procedure that derives subjective probabilities from preferences between bets. Holding the payoff constant, we conclude that the subject regards A as more probable than B whenever he or she prefers to bet on A rather than on B. Although this procedure seems reasonable, it is based on an assumption, called source independence, according to which preferences between risky prospects depend on the degree of uncertainty but not on its source. Thus, if the decision maker regards two propositions as equally likely, he or she should be equally willing to bet on either one. Indeed, this assumption underlies the modern theory of utility and subjective probability (Ramsey, 1931; Savage, 1954).

The assumption of source independence has been challenged by Ellsberg (1961) who showed that people prefer to bet on drawing a red ball from a box containing 50 red and 50 green balls than from a box that contains an unknown number of green and red balls, even though they have no color preferences. This pattern violates the additivity of subjective probability and suggests that people prefer to bet on known rather than on unknown chances. Although this pattern has been observed in the study of chance processes, it does not extend

readily to judgmental probabilities based on an assessment of evidence. In particular, people often prefer to bet on their skill rather on chance (Howell, 1971) even though the former is vaguer than the latter, and people seem to prefer betting on the future rather than on the past (Rothbard & Snyder, 1970) even though the future is less knowable than the past.

Chip Heath and I have recently developed a new account of uncertainty preferences, called the competence hypothesis, which applies to both chance and evidential problems. We submit that the willingness to bet on an uncertain event depends not only on the estimated likelihood of that event (subjective probability) and the precision of that estimation (ambiguity or vagueness), it also depends on one's general knowledge or understanding of the relevant context. More specifically, we propose that -- holding degree of belief or judged probability constant -- people prefer to bet in a context where they consider themselves knowledgeable or competent than in a context where they feel ignorant or uninformed. We assume that the feeling of competence in a given context is enhanced by general knowledge, familiarity and experience, and it is diminished, for example, by calling attention to relevant information that is not available to the decision maker, especially if it is available to others.

There are both judgmental and preferential reasons for the competence hypothesis. First, people may have learned from life-long experience that they generally do better in situations they understand or control than in situations in which they have less knowledge and competence. Thus, they expect to do better in the former case, and this feeling may carry over to situations where the chances of winning are no longer higher in the familiar than in the unfamiliar context. Second, people may like to bet on their (physical or mental) skills, either because they enjoy the challenge or because they like to demonstrate their competence. Conversely, people may avoid

a situation they do not understand because it makes them feel incompetent.

The competence hypothesis readily applies to Ellsberg's example. People do not like to bet on the unknown box, we suggest, because there exists relevant evidence, namely the proportion of red and green balls in the box, that is knowable in principle but unknown to them. The presence of such data make people feel less knowledgeable and less competent, and reduce the attractiveness of the corresponding bet. This account is also consistent with the finding of Curley, Yates, and Abrams (1986) that the aversion to ambiguity is enhanced by the anticipation that the content of the unknown box will be shown to others. Essentially the same analysis applies to the preference for betting on the future rather than on the past. Because the past, unlike the future, is known to others but not to themselves, subjects prefer to bet on the future where their relative ignorance is lower. In prediction, only the future can prove you wrong; in postdiction, you could be wrong right now. Recall that in the 50/50 box, a guess could turn out to be wrong only after drawing the ball. In the unknown box, on the other hand, the guess may turn out to be mistaken even before the drawing of the ball -- if it turns out that the majority of balls in the box are of the opposite color. The competence hypothesis can also explain the preference for betting on skill over chance, whenever the subjects regard themselves as skillful or competent.

The competence hypothesis is supported in a series of experiments showing that people prefer to bet on their belief over a matched chance event when they feel knowledgeable or competent, and they prefer to bet on chance events over their judgment when they feel ignorant or uninformed (Heath & Tversky, 1990). In the first experiment, subjects answered 30 knowledge questions in various categories, such as history, geography or sports. Four alternative answers were presented for each question, and the subjects first selected a single answer and then rated

his or her confidence in that answer on a scale from 25% (pure guessing) to 100% (absolute certainty). Participants were instructed to use the scale so that a confidence rating of 60%, say, would correspond to a hit rate of 60%. After answering the questions and assessing confidence, subjects were given an opportunity to choose between betting on their answer or on a lottery in which the probability of winning was equal to their stated confidence. For a confidence rating of 75%, for example, the subject was given a choice between (i) betting that his or her answer was correct, or (ii) betting on a 75% lottery, defined by drawing a number chip in the range of 1 to 75 from a bag filled with 100 numbered poker chips.

The dependent variable in the study was the percentage of responses that favored the judgment bet over the matched chance lottery. If source independence holds, subjects should be indifferent between the two bets. Under ambiguity aversion, subjects should prefer the chance lottery, that is precisely defined, over the judgment bet where the probability is generally ambiguous or vague. The competence hypothesis, on the other hand, predicts a preference for the judgment bet when the probability associated with the response is high and a preference for the chance lottery when the probability associated with the response is low. This is exactly the pattern observed in the study. More specifically, the percentage of responses that favored the judgment bet was a monotonic function of the judged probability. This result has been replicated in several settings, including the prediction of sport events and election results.

In the preceding experiment, competence, or knowledge, was confounded with judged probability. To provide a sharper test of the competence hypothesis, we have sorted subjects according to their expertise. To this end, we have selected a group of subjects who consider themselves experts in football but not in politics, and another group of subjects who consider

themselves experts in politics but not in football. In the first part of the experiment, each subject made predictions for 40 future events (20 political events and 20 football games). Political events concerned the winner of various states in the 1988 presidential election, the football games included 10 professional and 10 college games. For each contest (politics or football), subjects chose a winner and assessed the probability that their prediction would come through. Following the assessment stage, 20 triples of bets were constructed for each participant. Each triple included three matched bets with the same probability of winning generated by (i) a chance devise, (ii) the subject's prediction in his or her strong category, and (iii) the subject's prediction in his or her weak category. In the second session, subjects ranked each of the 20 triples of bets.

The results were very clear. At all probability levels, the high-knowledge bet was preferred to the chance bet which in turn was preferred to the low-knowledge bet. In other words, subjects preferred the judgment bet over the matched chance lottery in their area of expertise but they preferred the chance lottery in an area where they did not consider themselves experts. These results cannot be explained in terms of the accuracy of prediction. The expected proportion of winning was about 70% for the chance lottery, 65% for the weak category, and only 60% for the strong category. By betting on the expert category, therefore, the subjects were losing in effect 15% of their expected earning.

The experimental finding establishes a consistent and pervasive discrepancy between judgments of probability and a choice between bets, as implied by the competence hypothesis. This hypothesis can be used to explain other instances of uncertainty preferences reported in the literature, notably the preference for a clear over a vague probability in a chance setup (Ellsberg,

1961), the preference to bet on the future over the past (Rothbard & Snyder, 1970), and the preference for skill over chance (Howell, 1971). Our major empirical finding that, in their area of competence, people prefer to bet on their vague beliefs over a matched chanced event demonstrates that the effect of knowledge or competence far outweighs the contribution of ambiguity.

The present conclusion challenges not only expected utility theory; it also challenges the very idea of using preferences to infer beliefs. For if people's willingness to act depends not only on the degree of uncertainty (and the precision with which it is measured) but also on one's general knowledge of the domain and his or her sense of competence concerning a particular proposition, it is exceedingly difficult, if not impossible, to derive underlying beliefs from observed preferences.

References

Curley, S. P., Yates, J. F., & Abrams, R. A. (1986). Psychological sources of ambiguity avoidance. *Organizational Behavior and Human Decision Processes, 38,* 230-256.

Edwards, W. (1968). Conservatism in human information processing. In B. Kleinmuntz (Ed.), *Formal representation of human judgment* (pp. 17-52). New York: Wiley.

Ellsberg, D. (1961). Risk, ambiguity, and the Savage axioms. *Quarterly Journal of Economics, 75,* 643-669.

Heath, F., & Tversky, A. (1990). Ambiguity and confidence in choice under uncertainty. In D. Messick (Ed.), *Proceedings of the Twelfth Research Conference on Subjective Probability, Utility and Decision Making.*

Howell, W. C. (1971). Uncertainty from internal and external sources: A clear case of overconfidence. *Journal of Experimental Psychology, 89* (2), 240-243.

Kahneman, D., Slovic, P., & Tversky, A. (1982). *Judgment under uncertainty: Heuristics and biases.* Cambridge: Cambridge University Press.

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review, 80,* 237-251.

Lichtenstein, S., & Slovic, P. (1971). Reversals of preference between bids and choices in gambling decisions. *Journal of Experimental Psychology, 89,* 46-55.

Ramsey, F. P. (1931). Truth and probability. In F. P. Ramsey, *The foundations of mathematics and other logical essays.* NY: Harcourt, Brace and Co.

Rothbart, M., & Snyder, M. (1970). Confidence in the prediction and postdiction of an uncertain outcome. *Canadian Journal of Behavioral Science, 2* (1), 38-43.

Savage, L. J. (1954). *The foundations of statistics*. NY: Wiley.

Schkade, D. A., & Johnson, E. J. (1989). Cognitive processes in preference reversals. *Organization Behavior and Human Performance, 44,* 203-231.

Slovic, P., Fischhoff, B., & Lichtenstein, S. (1982). Response mode, framing, and information-processing effects in risk assessment. In R. Hogarth (Ed.), *New directions for methodology of social and behavioral science: Question framing and response consistency,* no. 11 (pp. 21-36). San Francisco, CA: Jossey-Bass.

Slovic, P., Griffin, D., & Tversky, A. (1990). Compatibility Effects in Judgment and Choice. In R. M. Hogarth (Ed.), *Insights in decision making: Theory and applications*. IL: University of Chicago Press.

Taylor, S. E., & Brown, J. D. (1988). Illusion and well-being: A social psychological perspective on mental health. *Psychological Bulletin, 103,* 193-210.

Tversky, A., & Griffin, D. (1990). *The weighting of evidence and the determinants of confidence*. Unpublished manuscript.

Tversky, A., & Kahneman, D. (1971). The belief in the "law of small numbers. *Psychological Bulletin, 76,* 105-110.

Tversky, A., Sattath, S., & Slovic, P. (1988). Contingent weighting in judgment and choice. *Psychological Review, 95* (3), 371-384.

Tversky, A., Slovic, P., & Kahneman, D. (in press). The Causes of Preference Reversal. *American Economic Review.*

Tversky, A., & Thaler, R. (in press). Preference reversals. *Journal of Economic Perspectives.*

The Weighing of Evidence and the Determinants of Confidence

Amos Tversky and Dale Griffin

Stanford University and The University of Waterloo

Abstract

Research on the intuitive weighting of evidence reveals that some components of evidence, notably extremeness or strength, dominate other components of evidence, such as credence or weight. We propose that the pattern of overconfidence and underconfidence observed in the literature is produced by the relations of dominance between components of evidence. In particular, overconfidence arises when the strength of evidence is high but the weight of evidence is low (e.g., an extreme proportion based on a small sample, or a glowing letter of recommendation by an uninformed source), and underconfidence occurs when strength is relatively low and weight is high. We first demonstrate the relation of dominance and its implications for judgments of confidence in a chance setup, and then investigate its application to more complex evidential problems. In particular, we show that people are generally overconfident in predicting the behavior of others, and are sometimes underconfident in predicting their own behavior.

The Weighing of Evidence and the Determinants of Confidence

Amos Tversky and Dale Griffin

Stanford University and The University of Waterloo

The weighing of evidence and the formation of beliefs are basic elements of human thought. The question of how to evaluate evidence and assess confidence has been addressed from a normative perspective by philosophers and statisticians; it has also been investigated experimentally by psychologists and economists. A major empirical generalization that has emerged from this research is the observation that both lay people and experts are typically more confident in their judgments than is warranted by the facts.

Overconfidence is manifest in action, as well as in belief. Not only do people express high confidence in judgments of low or moderate validity (Dawes, 1988; Kahneman, Slovic & Tversky, 1982; Slovic, Lichtenstein & Fischhoff, 1988), they also act and commit resources on the basis of inadequate evidence. The well-publicized observation that more than 85% of new small businesses fail within one year suggests that many entrepreneurs overestimate their probability of success. With the notable exception of weather forecasters (Murphy & Winkler, 1977), who receive immediate frequentistic feedback and produce realistic forecasts of precipitation, overconfidence has been observed in judgments of physicians (Lusted, 1977), clinical psychologists (Oskamp, 1965), lawyers (Malsch, 1988), engineers (Kidd, 1970), and security analysts (Staël von Holstein, 1972). As one critic characterized the performance of experts, "often wrong but rarely in doubt."

Overconfidence is pervasive but not universal. One context in which people seem to exhibit underconfidence is the prediction of their own action. When people say that the odds are 2 to 1 that they will take job A rather than job B, the actual odds appear more extreme. Over the past few years we have discreetly approached colleagues faced with a choice between job offers, and asked them to estimate the probability that they will choose one job over another. The average confidence in the predicted choice was a modest 66%, but only 1 of the 24 respondents chose the option to which he or she assigned a lower probability, yielding an overall accuracy rate of 96%. This observation is suggestive but not conclusive because the underconfidence observed in this context can be attributed to factors that govern the expression of confidence rather than the evaluation of evidence. People may enjoy the sense of freedom associated with the pre-decisional state, and they may be reluctant to appear overconfident. The following study was designed to investigate the presence of underconfidence in the prediction of one's own action.

Study 1: Overconfidence and underconfidence.

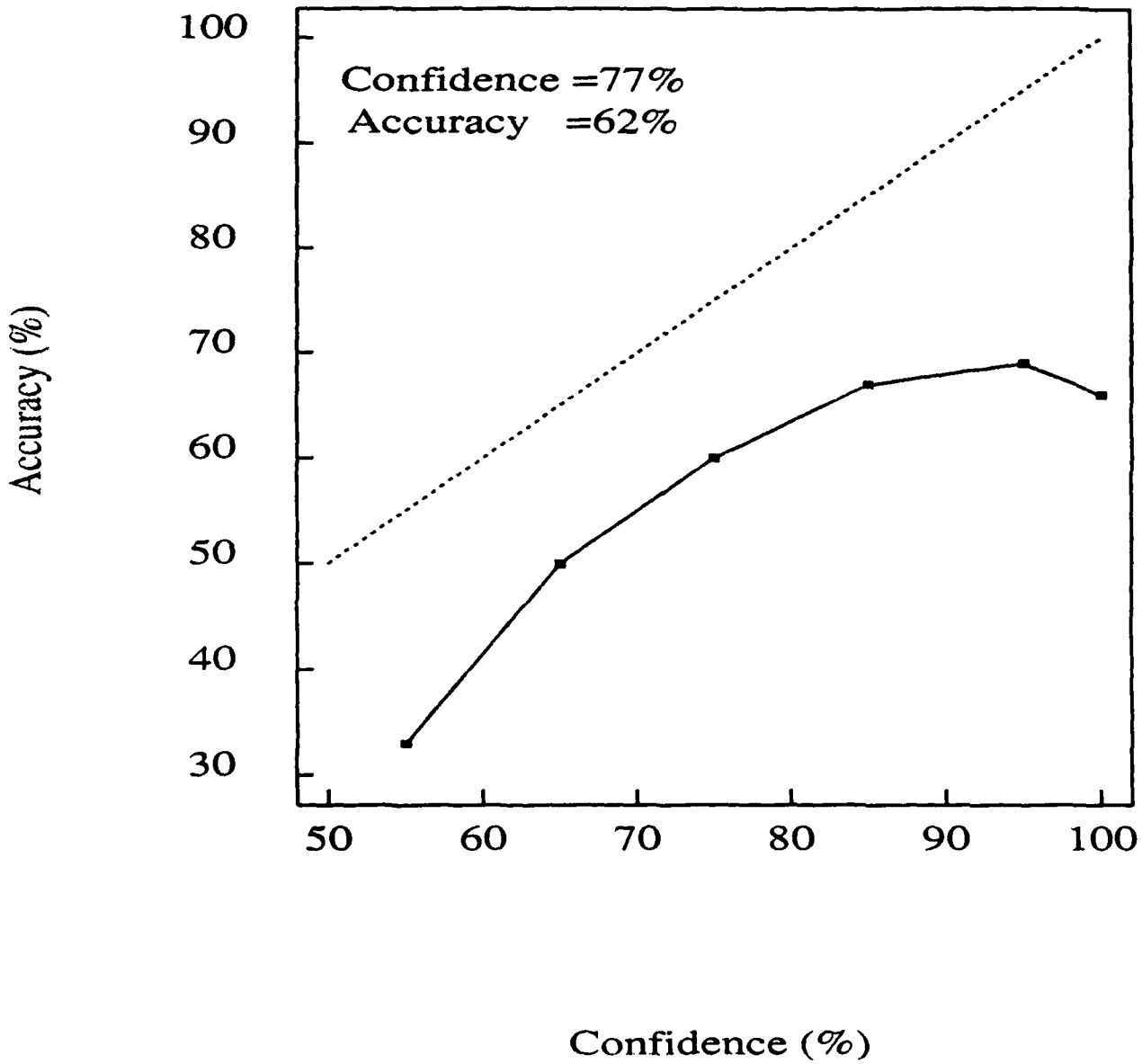
Overconfidence has been commonly demonstrated in the calibration paradigm in which the judge evaluates the probability of uncertain events, such as the possible answers to a multiple-choice test (Lichtenstein, Fischhoff & Phillips, 1982). As noted by several authors (see, e.g., May, 1988), the inclusion of "tricky" problems in which the intuitive response is incorrect, is likely to yield overconfidence, whereas the inclusion of "give-away" items is likely to yield underconfidence. To avoid bias in the selection of items, we chose twelve uncertain

events whose outcomes would become known in the two weeks following our study, such as "the San Francisco Forty-Niners will win or tie their next two games", or "the Cosby show will be the top rated television show by the Nielsen Company at least once in the next two weeks." A group of 111 undergraduates were asked to state their probability for each of the events. They were instructed to use the scale so that a 75% confidence, say, would correspond to a 75% hit rate. The results of the study are summarized in Figure 1.

Insert Figure 1 about here

Overall, subjects were overconfident: their average confidence across all twelve events was 77%, while the average accuracy was only 62%, ($p < .001$). Figure 1 is a calibration plot. If people were perfectly calibrated, then the solid line should coincide with the broken line. That is, 75% of the statements to which the judge assigned a chance of 75% should turn out to be correct. In contrast, the figure shows that barely 60% of statements to which the subjects assigned chances of 75% (or more accurately between 70% and 80%) turned out to be true. The finding that the solid line lies below the broken line indicates that peoples' confidence in the prediction of unknown future events exceeded their accuracy. In the binary problems used throughout this paper, a person is said to be overconfident if his or her probability judgments are more extreme (i.e., closer to 0 or 1) than is appropriate. The appropriate standard of comparison may be determined empirically (by a person's hit rate) or theoretically (by reference to the normative theory).

Figure 1

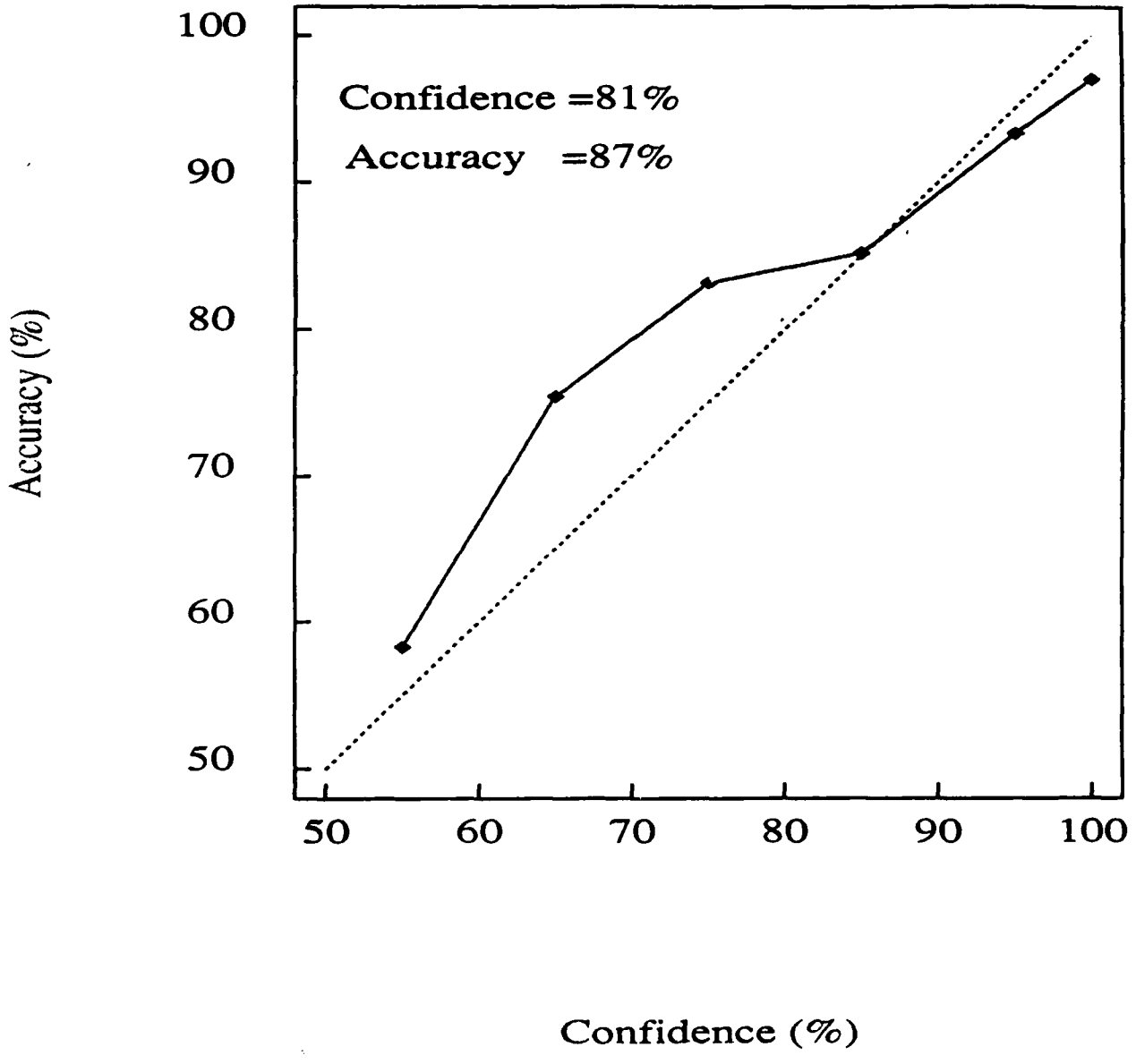

Confidence =77%
Accuracy =62%
100
90
80
70
60
50
40
30
50
60
70
80
90
100
Accuracy (%)
Confidence (%)

To test the calibration of predictions of one's own responses, we presented a comparable group of 60 Stanford undergraduates with a mock poll concerning social and political attitudes. The poll consisted of pairs of statements such as (a) the price of gas should be increased to encourage energy conservation vs. (b) the price of gas should continue to be set by market forces, or (a) Stanford should broaden its Western Civilization requirements to include more non-western themes vs. (b) Stanford should continue to focus on the Classical influence on western culture. Subjects were asked to state the probability that they would select one statement over another if they were presented with such a poll sometime in the course of the semester. A couple of months later, the same subjects were contacted again and were presented with the poll described earlier. By comparing their actual answers to their estimated probabilities, it is possible to plot the calibration curve described in Figure 2.

Insert Figure 2 about here

These data differ markedly from those displayed in Figure 1. In contrast to the pronounced overconfidence observed in the prediction of future events, the respondents' predictions of their own behavior were underconfident. Overall, average confidence was 81% while average accuracy was 87% ($p < .01$). Inspection of the graph indicates that except at the very high end, where there is little room for underconfidence, much of the solid line lies above the broken line. Because the same response scale was used in the two studies, neither overconfidence nor

Figure 2


100
90
80
70
60
50
Accuracy (%)
Confidence =81%
Accuracy =87%
50
60
70
80
90
100
Confidence (%)

underconfidence can be explained by a systematic bias in the use of the scale.[1]

A related pattern of overconfidence and underconfidence has been observed in other experimental paradigms. Early studies (e.g., Edwards, 1968) concluded that judgments of posterior probability are generally conservative (i.e., underconfident) in the sense that they are less extreme than warranted by the available data. However, later research (e.g., Tversky & Kahneman, 1971) showed that people are often overconfident in inferences based on small samples. In a different research paradigm, Kunda and Nisbett (1986) concluded that people are generally overconfident in predicting the results of small samples and are often underconfident in predicting the results of large samples. These findings raise the general question of what are the factors that give rise to overconfidence or to underconfidence? The present article is devoted to the analysis of this question.

The Weighing of Evidence

Confidence is important because it controls action. People generally act on beliefs that are held with a high degree of confidence and are reluctant to act in the presence of doubt (Tversky & Heath, 1989). Either overconfidence or underconfidence, therefore, can lead to inappropriate action. The major thesis of the present article is that the observed pattern of overconfidence

---

[1] It is noteworthy that the overconfidence observed in calibration studies greatly diminishes or disappears when people are asked to estimate their overall hit rate rather than their confidence in a specific question (e.g., Dunning, Milojkovic, Griffin, & Ross, in press; Kleinbolting & Gigerenzer, 1989). Evidently, people can maintain a high degree of confidence in the validity of specific answers even when they know that their overall hit rate is not very high. This is the probabilistic version of the paradoxical statement "I believe in all of my beliefs, but I believe that some of my beliefs are false".

and underconfidence is produced by an improper weighting of evidential components. Judgment under uncertainty typically requires the integration of different items of evidence. Statistical theory and the calculus of probability specify the rules for combining the components of evidence. The study of human judgment has shown that intuitive weighting of evidence departs systematically from these rules. In particular, some variables, such as the strength of evidence (sample proportion or the warmth of a letter of recommendation), tend to dominate other variables, such as the weight of evidence (sample size or the credibility of the writer). We say that variable A dominates variable B if A is consistently overweighted relative to B, in comparison to the appropriate normative theory.

A review of the judgment literature reveals several dominance relations between evidential variables (e.g., specific evidence or case data generally dominate statistical evidence or base rate data). These findings have been attributed to the operation of judgmental heuristic such as representativeness, availability, or anchoring (Kahneman, Slovic, & Tversky, 1982), as well as to attentional factors such as vividness or salience (Fiske & Taylor, 1984). The present paper does not address the explanation of dominance relations. Instead, we argue that both overconfidence and underconfidence follow from the pattern of dominance between components of evidence. More specifically, we propose the following hypotheses.

(a) The strength or extremeness of evidence generally dominates its weight or credence. For example, the subjective probability that a particular experimental result will be replicated is highly sensitive to the size of the effect (e.g., the difference between the means), and not sufficiently sensitive to sample size, which reflects its weight or credence (Tversky & Kahneman, 1971). If people are highly sensitive to the strength of evidence and not

sufficiently sensitive to its weight, then their judgments will be overconfident when strength is high and weight is low, and their judgments will be underconfident when weight is high and strength is low.

(b) Although reliability information is dominated by strength, confidence in a prediction is more sensitive to the reliability of the predictor than to the reliability of the criterion. As a consequence, people are more willing to predict from a reliable variable (e.g., a student's GPA) to a less reliable criterion (e.g., a grade on a single exam) than vice versa (Tversky & Kahneman, 1980). Therefore overconfidence is expected in the prediction of an unreliable criterion from a reliable predictor and underconfidence is expected in the prediction of a reliable criterion from an unreliable predictor (e.g., predicting the result of an election from a random sample).

(c) Specific evidence or case data are generally weighted more heavily than statistical evidence or base-rate data. For example, people predict a person's occupation on the basis of limited personal information (e.g., avocation or a personality test) with insufficient regard for the base rate frequency of the various occupations (Tversky & Kahneman, 1973). Consequently, overconfidence is expected when the base rate is low and underconfidence is expected when the base rate is high.

(d) The degree to which the evidence fits the focal hypothesis has more impact on intuitive judgments than the fit between the evidence and an alternate (non-focal) hypothesis. People often accept a focal hypothesis on the basis of confirming data without testing whether the same data are equally compatible with an alternate hypothesis (Fischhoff & Beyth-Marom, 1983). Overconfidence in the validity of the focal hypothesis, therefore, is expected when the evidence is compatible with both the focal hypothesis and its alternative.

Underconfidence is expected when the evidence does not fit the focal hypothesis very well, but the fit to the alternate hypothesis is even poorer.

We first test these hypotheses in a chance setup, where the components of evidence are well-defined and their weights are determined by probability theory. We then extend the analysis to more complicated evidential problems for which the appropriate normative model is not fully specified. In each case, we first establish the relation of dominance between the relevant variables and then demonstrate how it gives rise to overconfidence and underconfidence. It is important to note that the present analysis does not specify the exact boundaries of overconfidence and underconfidence; it merely defines the direction of departure from the normative model. This account should also be distinguished from the hypothesis that people are not sufficiently sensitive to variations in all relevant evidential variables. In contrast, we propose that people are overly sensitive to variations in some variables and relatively insensitive to variations in others.

Evaluating Evidence

Study 2: Strength versus weight

We first investigate the relative impact of strength and weight in an experimental task involving the assessment of posterior probability. We presented 16 Stanford students with the following instructions:

> "Imagine that you are spinning a coin, and recording how often the coin lands heads and how often the coin lands tails. Unlike tossing, which (on average) yields an equal number of heads and tails, spinning a coin leads to a bias favoring one side or the other because of slight imperfections on the rim of the coin (and an uneven distribution of mass). Now imagine that you know that this bias is 3/5.

> It tends to land on one side 3 out of 5 times. But you do not know if this bias is in favor of heads or in favor of tails."

Subjects were then given various samples of evidence differing in sample size (from 3 to 33) and in of heads (from 2 to 19). All samples contained a majority of heads, and subjects were asked to estimate the probability (from .5 to 1) that the bias favored heads (H) rather than tails (T). Subjects received all 12 combinations of sample proportion and sample size shown in Table 1. They were offered a prize of $20 for the person whose judgments most closely matched the correct values.

Insert Table 1 about here

Table 1 also presents, for each sample of data (D), the posterior probability for hypothesis H (a 3:2 bias in favor of heads) computed according to Bayes' Rule. Assuming equal prior probabilities, Bayes' Rule yields

$$\log\left[\frac{P(H|D)}{P(T|D)}\right] = n\left[\frac{h}{n} - \frac{t}{n}\right]\log\left[\frac{.6}{.4}\right],$$

where h and t are the number of heads and tails, respectively, and n = h + t denotes the sample size. The first term on the right-hand side, n, represents the weight of evidence. The second term, which is the difference between the proportion of heads and tails in the sample, represents the strength of the evidence for H against T. The third term, which is held constant in this study, is the discriminability of the two hypotheses, which corresponds to d′ in signal detection theory. Plotting equal-support lines for strength and weight in logarithmic coordinates yields a family of parallel straight lines with a slope of -1, as illustrated by the dotted lines in Figure 3. (To facili-

Table 1

Stimuli and Responses for Study 2

Number of Heads (h)	Number of Tail (t)	Sample Size (n)	Posterior Probability P(H|D)	Median Confidence (in %)
2	1	3	.60	63.0
3	0	3	.77	85.0
3	2	5	.60	60.0
4	1	5	.77	80.0
5	0	5	.88	92.5
5	4	9	.60	55.0
6	3	9	.77	66.9
7	2	9	.88	77.0
9	8	17	.60	54.5
10	7	17	.77	59.5
11	6	17	.88	64.5
19	14	33	.88	60.0

tate interpretation, the strength dimension is labeled $\frac{h}{n}$, which is a linear transformation of $\frac{h-t}{n}$). Each line connects all data sets that provide the same support for hypothesis H.

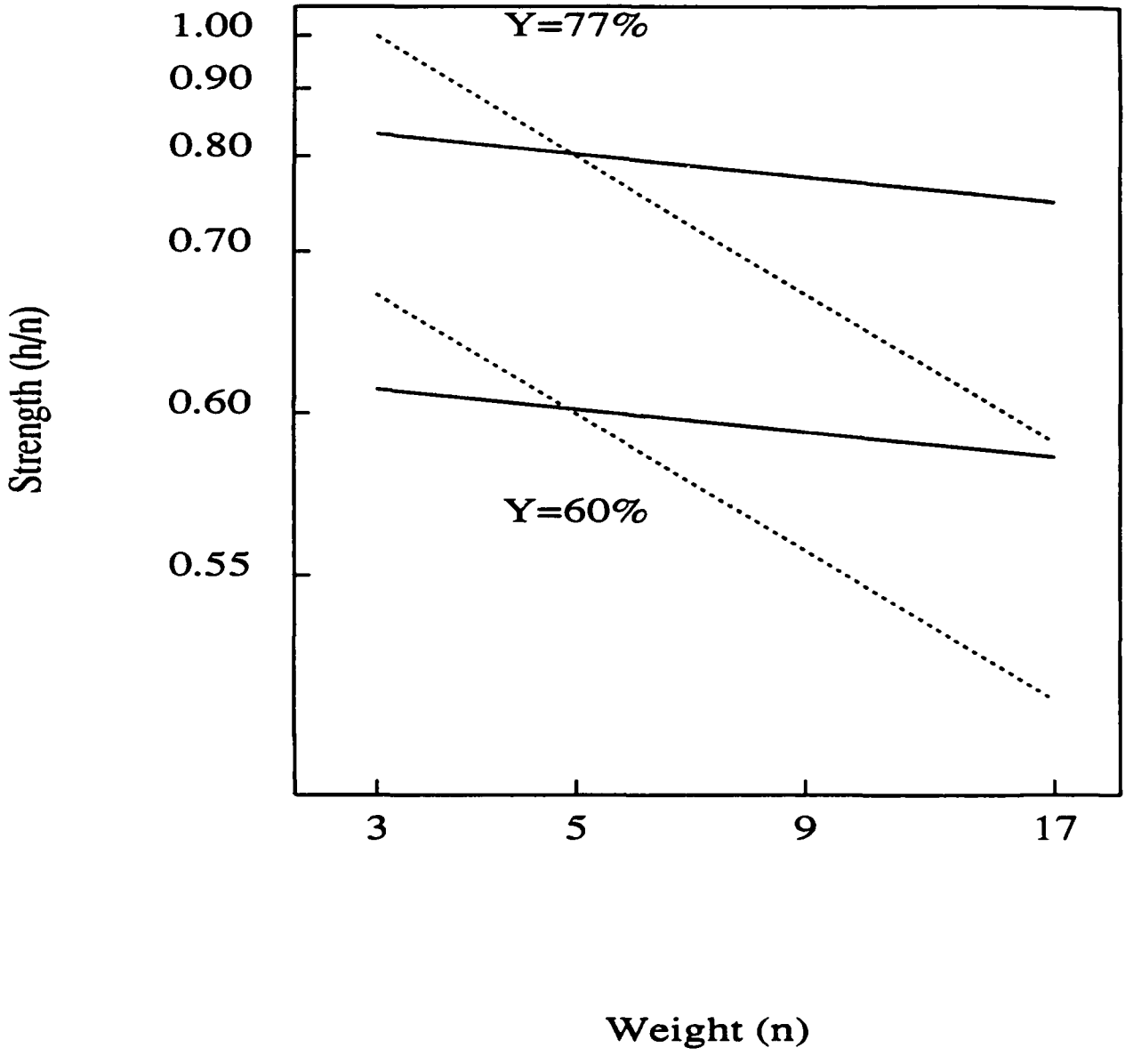
Insert Figure 3 about here

For example, a sample of size 9 with 6 heads and 3 tails, and a sample of size 17 with 10 heads and 7 tails, yields the same posterior probability (.77) for H over T. Thus the point (9, 6/9) and the point (17, 10/17) both lie on the upper line. Similarly, the lower line connects the data sets that yield a posterior probability of .60 in favor of H (see Table 1).

To compare the observed judgments with Bayes' Rule, we transformed each probability judgment into log odds and regressed the logarithm of the transformed values against the logarithm of strength, $\frac{h-t}{n}$, and weight, n. The regression fit the data quite well: multiple R was .98 for the median data and .92 for the median subject. According to Bayes' Rule, the regression weights for strength and weight in this metric are equal (see Figure 3). In fact, the observed regression weight for strength was nearly 6 times larger than that for weight (the actual coefficients were 1.0 and .17 for strength and weight, respectively).

The equal support lines obtained from the regression analysis are plotted in Figure 3 as solid lines. The comparison of the two sets of lines reveals two noteworthy findings. First, the intuitive lines are much shallower than the Bayesian lines indicating that the strength of evidence dominates its weight. Second, for a given level of support (e.g., 60% or 77%), the Baye-

Figure 3



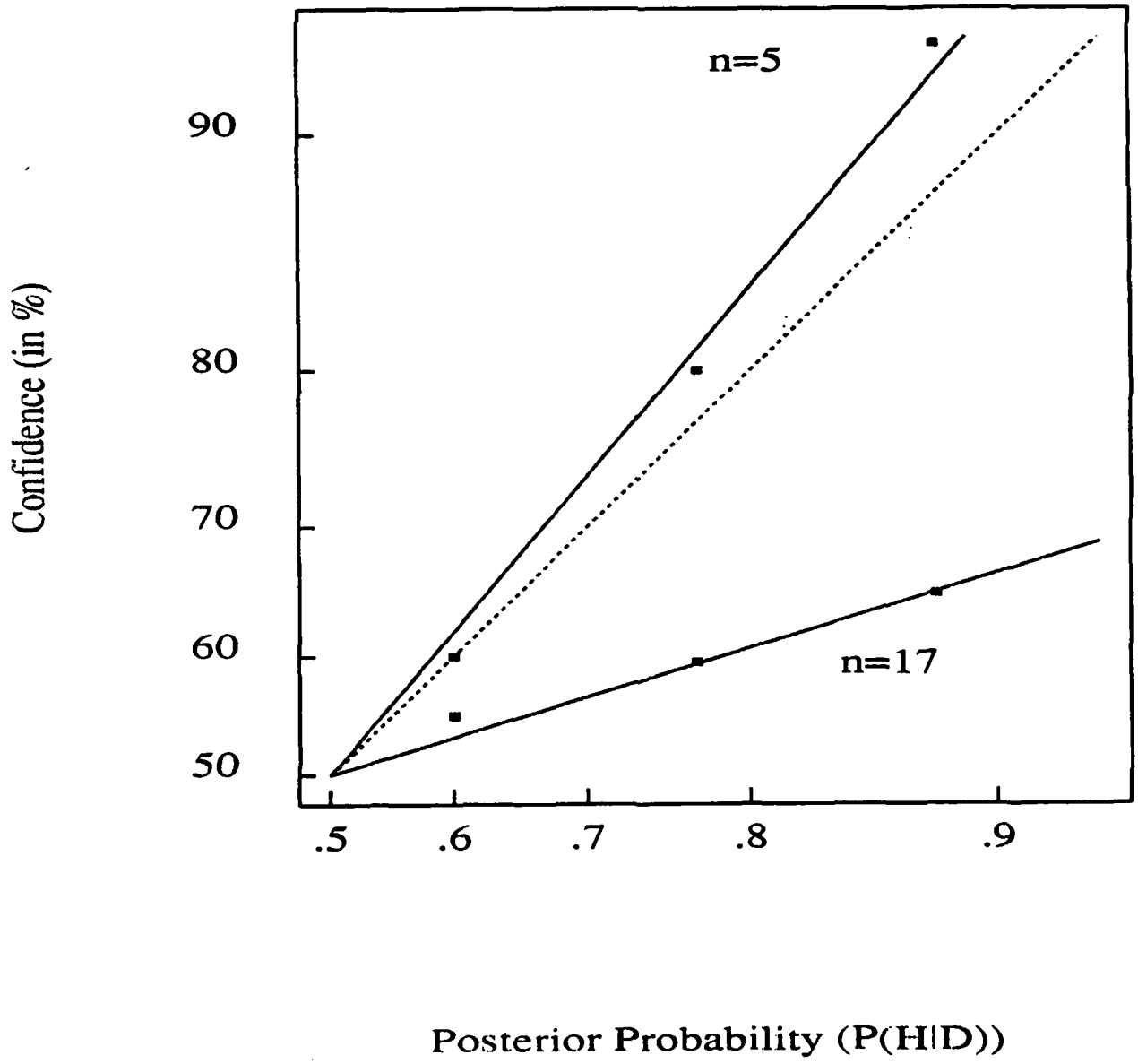
sian and the intuitive lines cross, indicating overconfidence where strength is high and weight is low, and underconfidence where strength is low and weight is high. As will be seen later, the crossing point is determined primarily by the discriminability of the competing hypotheses (d′).

Insert Figure 4 about here

Figure 4 plots the median confidence for a given sample of evidence as a function of the (Bayesian) posterior probability for two separate sample sizes. The best-fitting lines were calculated using the log odds metric. If the subjects were Bayesian, the squares should fall on the dotted line. Instead, intuitive judgments based on the small sample (n=5) were slightly overconfident, whereas the judgments based on the larger sample (n=17) were markedly underconfident.

The results described in Table 1 are in general agreement with previous results (see e.g., Kahneman, Slovic & Tversky, 1982; von Winterfeldt & Edwards, 1986). Moreover, they help reconcile apparently inconsistent findings. Edwards and his colleagues (e.g., Edwards, 1968) who used a sequential updating paradigm argued that people are conservative in the sense that they do not extract enough information from large samples of evidence. On the other hand, Tversky & Kahneman (1971), who investigated the role of sample size in researchers' confidence in the replicability of their results, concluded that people (even those trained in statistics) make radical inferences on the basis of small samples. Both results have been replicated in several studies. Figures 3 and 4 suggest how the dominance of sample proportion over sample

Figure 4


n=5
n=17
90
80
70
60
50
.5
.6
.7
.8
.9
Confidence (in %)
Posterior Probability (P(H|D))

size is responsible for both findings. In the updating experiments conducted by Edwards, subjects are exposed to large samples of data typically of moderate strength. This is the context in which we expect underconfidence or conservatism. The situations studied by Tversky and Kahneman, on the other hand, involve moderately strong effects based on fairly small samples. This is the context in which overconfidence is likely to prevail. Both conservatism and overconfidence, therefore, are generated by a common bias in the weighting of evidence, namely the dominance of strength over weight.

Study 3: Predictor versus criterion

Although the strength of evidence generally dominates its reliability, the reliability of the predictor has more impact than the reliability of the criterion. People are more willing to predict a fallible measure from a highly reliable measure than vice versa (Tversky & Kahneman, 1980). This hypothesis implies that people should be more confident in predicting from a population to a sample than from a sample to a population, even when the two predictions are normatively equivalent. To test this hypothesis, we asked eighty-five Stanford students from a class on decision making to indicate which of two bets they would rather play.

Bet A: A class of 100 students contains a majority of women (men). You win $10 if a random sample of 10 students taken from this class contains a majority of women (men).

Bet B: A random sample of 10 students from a class contained a majority of women (men). You win $10 if the class of 100 students contains a majority of women (men).

Half of the subjects received the version referring to a majority of women, and the other half received the version referring to a majority of men. Because there were no differences in the results, the data from the two groups were combined. Assuming that the chances of having a

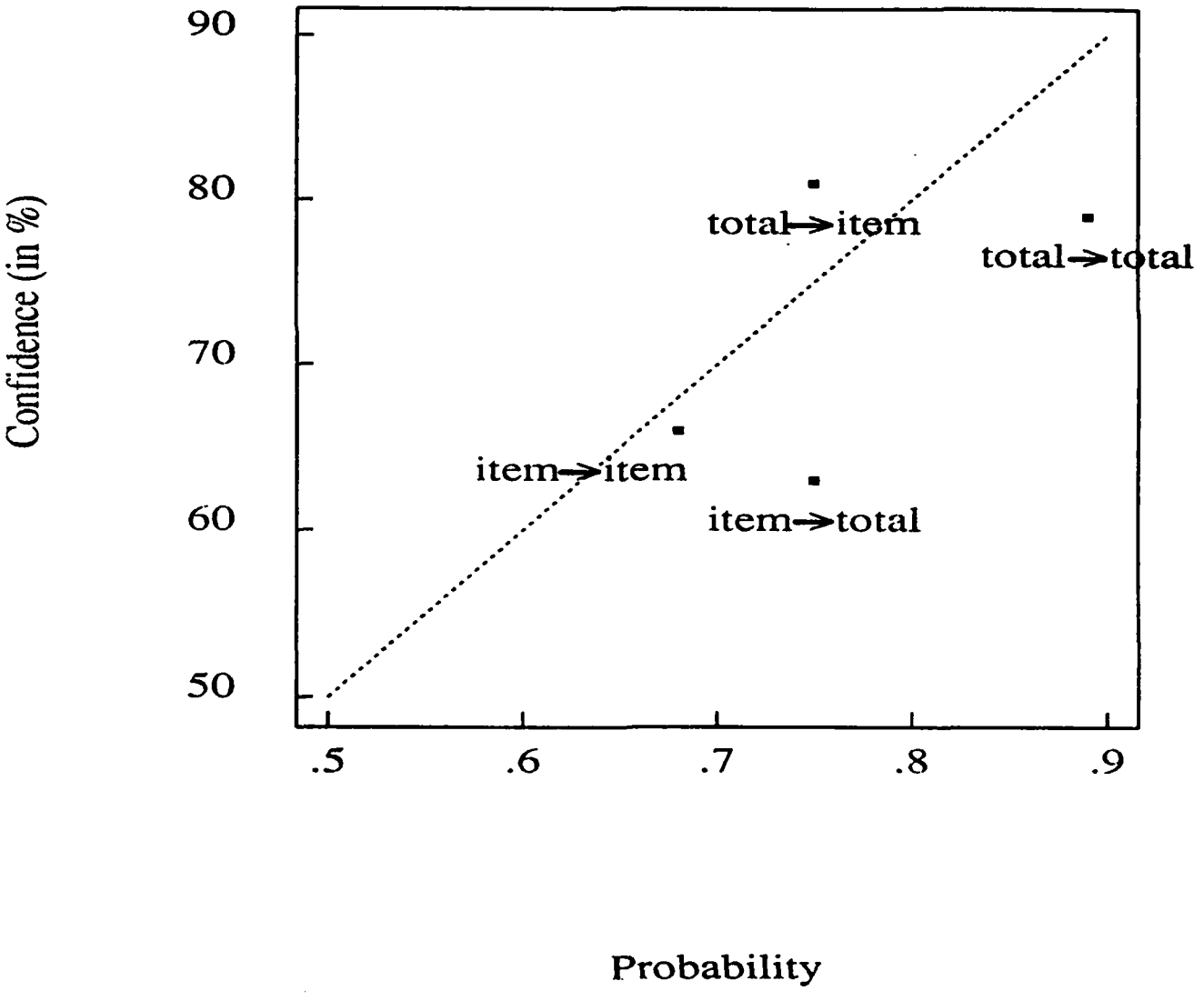
majority of women (men) in either the class or the sample is 50%, it follows readily that the probability of winning is the same in both bets. Nevertheless, 72% preferred Bet A involving the prediction of a sample from a population over Bet B involving the prediction of a population from a sample ($p<.05$).

Another study of the relation between the reliability of the predictor and the reliability of the criterion was undertaken by Kunda & Nisbett (1986). They told their subjects "In Jefferson school, students are required to take spelling tests each week in the 21 week term. Suppose you knew that Johnny got a higher grade than Danny on one (twenty) such test(s). What do you suppose is the probability that Johnny would get a higher grade than Danny's again a few weeks later, on the last (twenty) test(s) taken that term?" (p. 343). The authors first obtained an estimate of the item-item probability and computed the other three probabilities (item-total, total-item, and total-total) using the standard psychometric model. Figure 5 plots the average confidence judgment for all 4 conditions as a function of the computed (objective) probability.

Insert Figure 5 about here

As Figure 5 shows, subjects were more confident in predicting from a total score than in predicting from a single item, but they were no more confident when predicting the total score than when predicting a single item. Overall, the majority of subjects (69%) correctly expected greater predictability for larger predictor samples, but only a small minority of subjects (19%) expected greater predictability for larger criterion samples. This insensitivity to changes in the reliability of the criterion led to marked underconfidence in predicting the total score.

Figure 5


90
80
70
60
50
Confidence (in %)
total→item
total→total
item→item
item→total
.5
.6
.7
.8
.9
Probability

Insensitivity to the reliability of the criterion leads to overconfidence when the criterion is unreliable and the predictor is highly reliable (e.g., item-total judgments in Figure 5). This analysis can help explain apparently divergent findings regarding personality predictions. On the basis of extensive observations of a person's behavior, people often feel confident in predicting behavior in a single situation, leading to marked overconfidence (Dunning et. al, in press). Such overconfidence, however, is likely to diminish or disappear when people are asked to predict a highly aggregated outcome from a single behavioral observation (Kunda & Nisbett, 1986, Study 2).

Study 4: Individual versus base rate data

Considerable research has demonstrated that people tend to neglect background data (e.g. base rates) in the presence of specific evidence (Kahneman, Slovic & Tversky, 1982). This neglect can lead either to underconfidence or overconfidence, as will be shown below. We asked 40 Stanford students to imagine that they had three different foreign coins, each with a known bias of 3:2. As in Study 2, subjects did not know if the bias of each coin was in favor of heads (H) or in favor of tails (T). The subjects' prior probabilities of the two hypotheses (H and T) were varied. For one half of the subjects, the probability of H was .50 for one type of coin, .67 for a second type of coin and .90 for a third type of coin. For the other half of the subjects, the prior probabilities of H were .50, .33 and .10. Subjects were presented with samples of size 10, which included from 5 to 9 heads. They were then asked to give their confidence that the coin under consideration was biased in favor of heads.

Insert Table 2 about here

Again, a $20 prize was offered for the person whose judgments most closely matched the correct values. Table 2 summarizes the sample data, the posterior probability for each sample, and subjects' median confidence judgments. It is clear that our subjects overweighted strength of evidence and underweighted the prior probability.

According to Bayes Rule,

$$\log\left[\frac{P(H|D)}{P(T|D)}\right] = \log\left[\frac{P(H)}{P(T)}\right] + (h-t)\log\left[\frac{.6}{.4}\right].$$

In this metric, the posterior log odds are a linear function of prior odds and the discrepancy between heads and tails. The equal-support lines for strength and prior probability are plotted as dotted lines in Figure 6. The upper line represents those combinations of specific evidence (h-t) and prior probability $\frac{P(H)}{P(T)}$ that yield 90% support for hypothesis H. The lower line represents those combinations that yield 67% support for H.

Insert Figure 6 about here

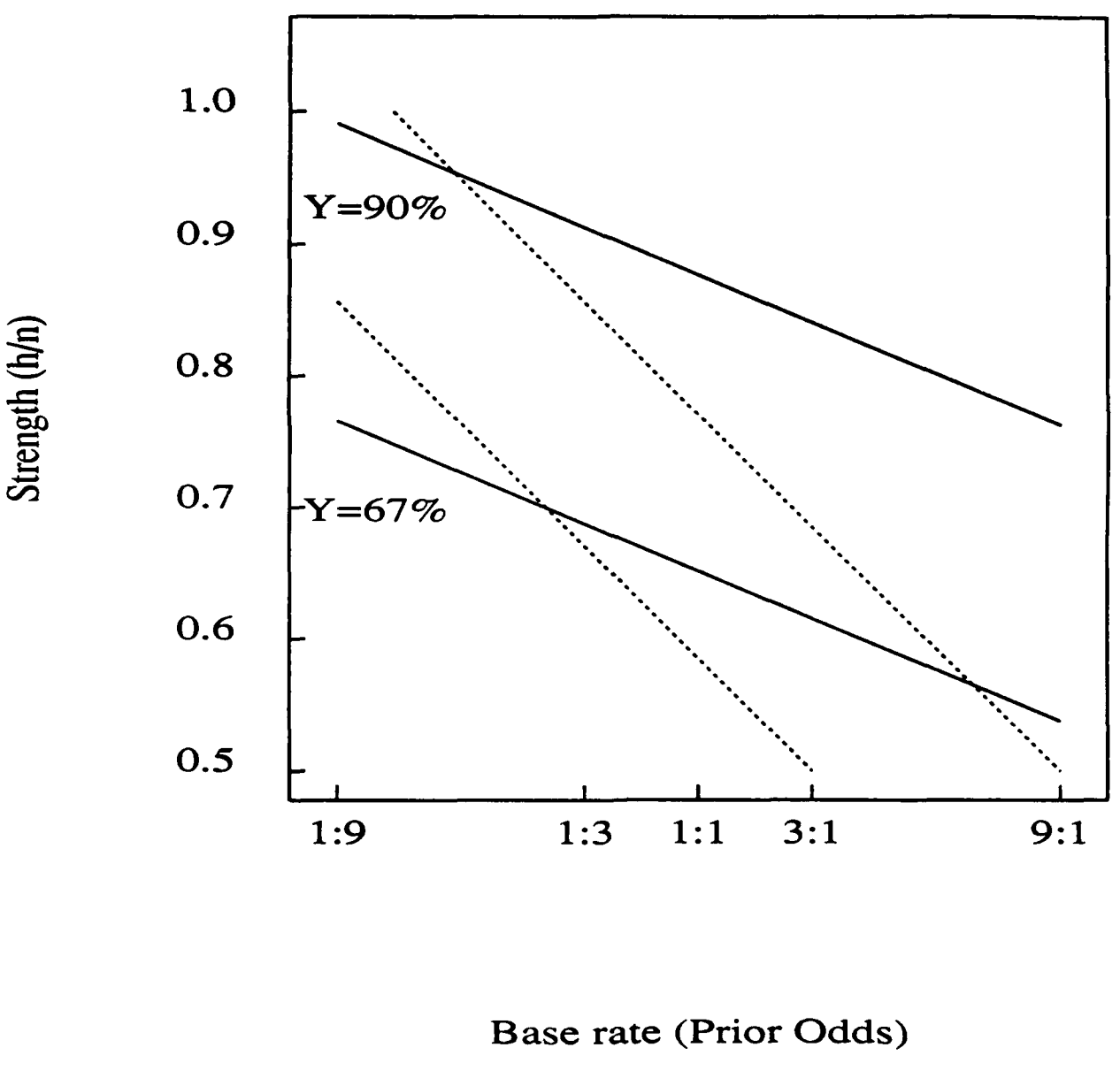
In order to compare our subjects' intuitive judgments with the Bayesian model, we regressed their median confidence judgments (in the form of log posterior odds) onto the

Table 2

Stimuli and Responses for Study 4

Number of Heads (out of 10)	Prior Probability (Base rate)	Posterior Probability P(H|D)	Median Confidence (in %)
5	9:1	.90	60.0
6	9:1	.95	70.0
7	9:1	.98	85.0
8	9:1	.99	92.5
9	9:1	.996	98.5
5	2:1	.67	55.0
6	2:1	.82	65.0
7	2:1	.91	71.0
8	2:1	.96	82.5
9	2:1	.98	90.0
5	1:1	.50	50.0
6	1:1	.69	60.0
7	1:1	.84	70.0
8	1:1	.92	80.0
9	1:1	.96	90.0
5	1:2	.33	33.0
6	1:2	.53	50.0
7	1:2	.72	57.0
8	1:2	.85	77.0
9	1:2	.93	90.0
5	1:9	.10	22.5
6	1:9	.20	45.0
7	1:9	.36	60.0
8	1:9	.55	80.0
9	1:9	.74	85.0

Figure 6


Strength (h/n)
1.0
0.9
0.8
0.7
0.6
0.5
Y=90%
Y=67%
1:9
1:3
1:1
3:1
9:1
Base rate (Prior Odds)

strength of evidence, (h-t) $\log\left[\frac{.6}{.4}\right]$, and prior probability , $\log\left[\frac{P(H)}{P(T)}\right]$. The regression model fit the median data very well, multiple R= .96 for the aggregate data, and .94 for the median subject. Recall that in the normative model the two weights are equal. In contrast, the best-fitting line for intuitive judgments yielded a coefficient of .82 for strength and .35 for prior probability, a ratio of 2.34 in favor of strength.

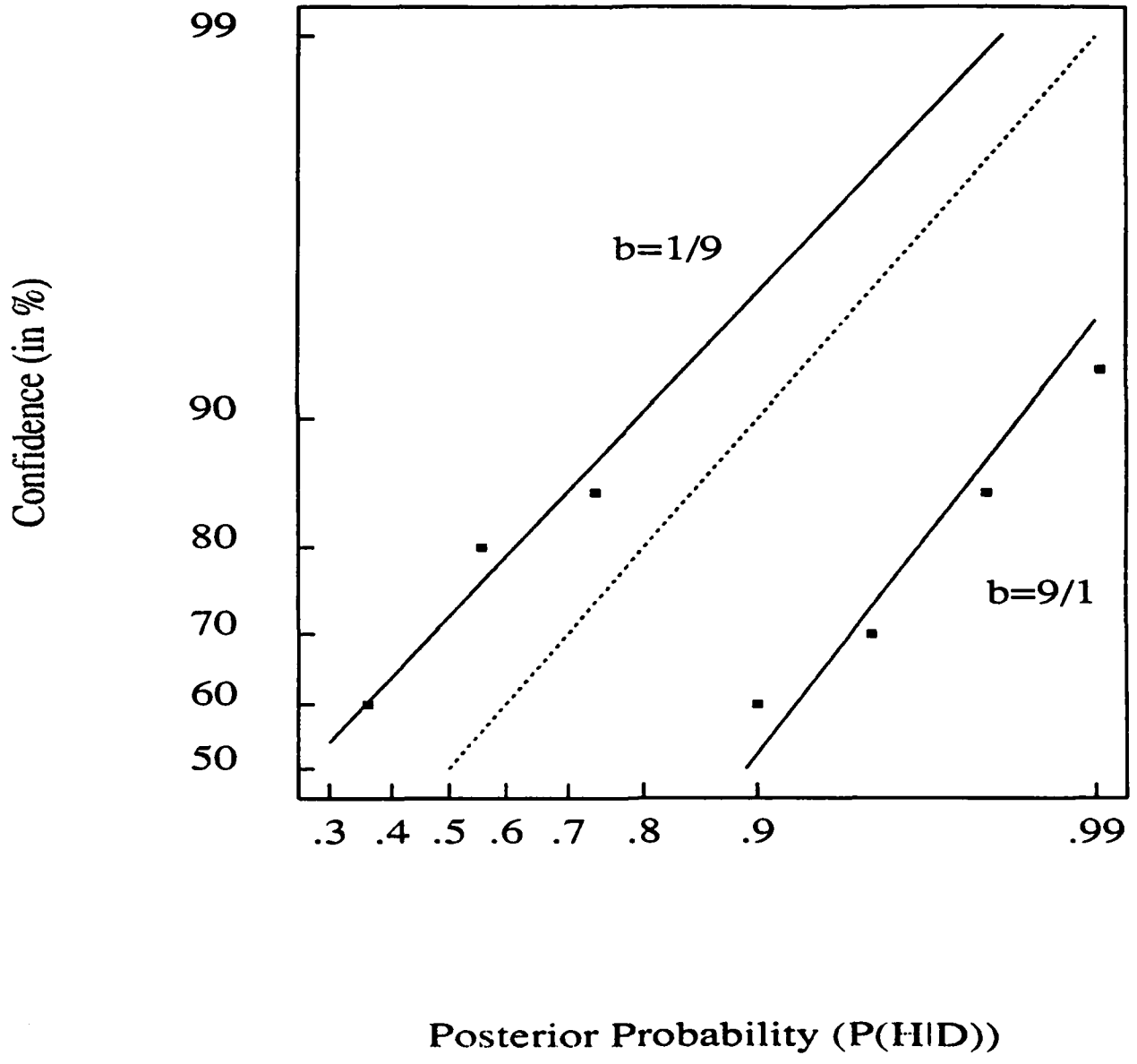
The solid lines in Figure 6 display the relative contribution of sample proportion and prior probability to the judged probability of H. As implied by the regression weights, the intuitive support lines are shallower than the normative support lines. Because the normative and intuitive support lines cross, judgments on the left side of the graph, where strength is high and prior probability is low, exhibit overconfidence, and judgments on the right side of the graph, where strength is low and prior probability is high, exhibit underconfidence (see Table 2).

Insert Figure 7 about here

Figure 7 plots median judgments of confidence as a function of (Bayesian) posterior probability for high (.90) and low (.10) prior probabilities of H. The figure also displays the best-fitting lines for each of the conditions. It is evident from the figure that subjects were overconfident in the low base rate condition and underconfident in the high base rate condition.

These results are consistent with Grether's (1980) study of the role of the representativeness heuristic in a Bayesian estimation task. He found that both experts and non-experts, with

Figure 7


99
90
80
70
60
50
Confidence (in %)
b=1/9
b=9/1
.3
.4
.5
.6
.7
.8
.9
.99
Posterior Probability (P(H|D))

and without financial incentives, overweighted the likelihood ratio relative to prior probability. In his experiments, as in ours, subjects did not ignore base rates altogether, but they did not give them sufficient weight. Insensitivity to base rate data in the context of social prediction has been demonstrated by Dunning et. al (in press)

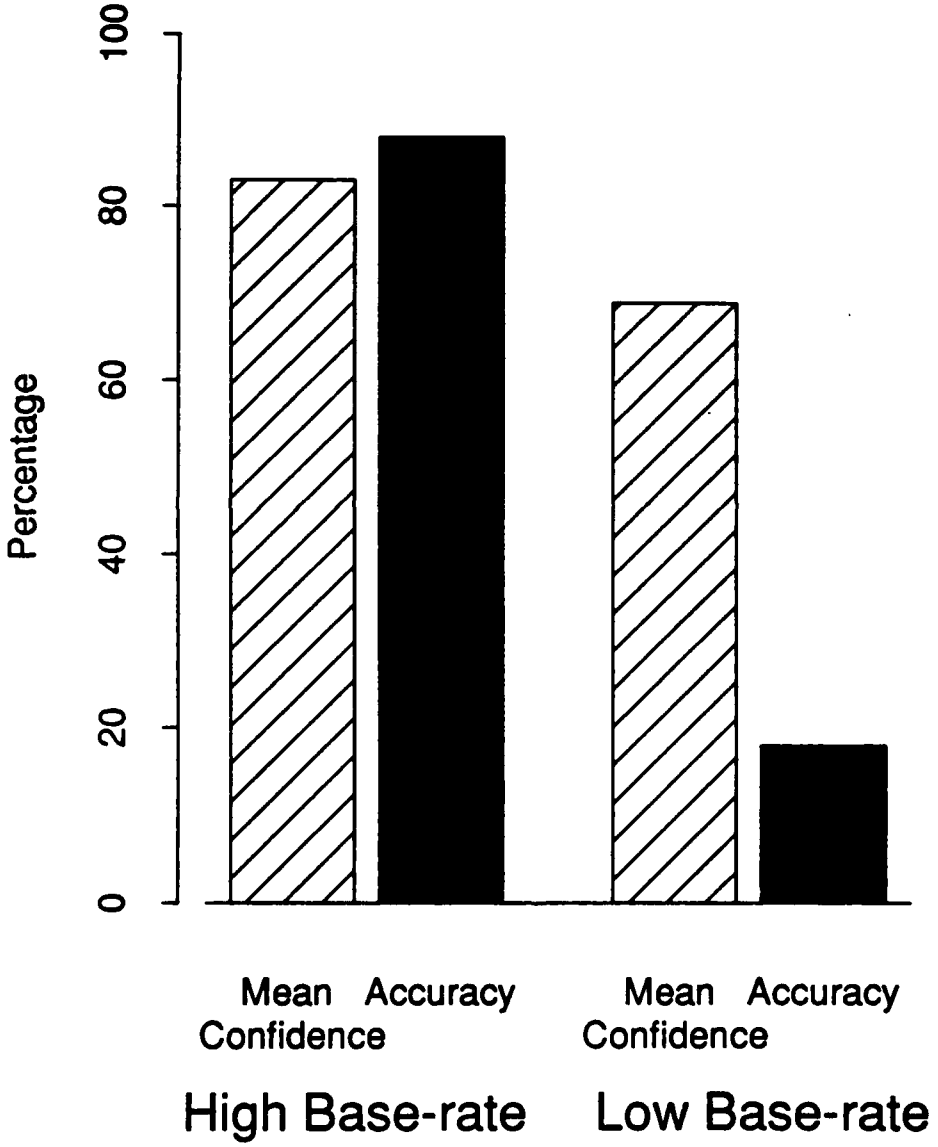
who asked subjects to assess their confidence in predictions about the behavior of a target person they had just interviewed. The predictions included such questions as "If this person were offered a free subscription, which magazine would he choose: Playboy or the New York Review of Books?" The authors also presented the subjects with empirically-derived estimates of the base rate frequency of the responses in question (e.g., that 68% of prior respondents preferred Playboy). In accord with the present analysis, Figure 8 shows that for the high base rate items (responses with prior odds of at least 3 to 1) subjects exhibited slight underconfidence, whereas for the low base rate items (responses with prior odds of at most 1 to 3) subjects were massively overconfident.

Insert Figure 8 about here

Study 5: Focal versus alternate hypothesis

When we consider the question of which of two hypotheses is true, confidence should depend on the degree to which the data fit one hypothesis better than the other. However, people seem to focus on the strength of evidence for a given hypothesis and neglect how well the same evidence fits an alternate hypothesis. The Barnum effect is a case in point. It is easy to

Figure 8


Percentage
0
20
40
60
80
100
Mean Confidence
Accuracy
Mean Confidence
Accuracy
High Base-rate
Low Base-rate

construct a personality sketch that will impress many people as a fairly accurate description of their own characteristics because they evaluate the description by the degree to which it fits their personality with little or no concern for whether it fits others just as well (Forer, 1949). To illustrate this effect in a chance setup, we presented 50 Stanford students with evidence about two types of foreign coins. Within each type of coin, the strength of evidence (sample proportion) varied from 7/12 heads to 10/12 heads. The two types of coins differed in their characteristic biases. Subjects were instructed:

> "Imagine that you are spinning a foreign coin called a *quinta*. Data have shown that half of the quintas (the "X" type) have a .6 bias towards Heads (that is, Heads comes up on 60% of the spins for X-quintas) and half of the quintas (the "Y" type) have a .75 bias toward Tails (that is, Tails comes up on 75% of the spins for Y-quintas). Your job is to determine if this is an X-quinta or a Y-quinta".

They then received the samples of evidence displayed in Table 3.

Insert Table 3 about here

After they gave their confidence that each sample came from an X-quinta or a Y-quinta, subjects were asked to make the same judgments for A-libnars (which have a .6 bias towards heads) and B-libnars (which have an .5 chance of heads). The order of presentation of coins was counterbalanced.

Table 3 summarizes the sample data, the posterior probability for each sample, and subjects' median confidence judgments. The comparison of the confidence judgments to the Bayesian posterior probabilities indicates that our subjects focused primarily on the degree to which the data fit the focal hypothesis with little or no regard for how well they fit the alternate

Table 3

Stimuli and Responses for Study 5

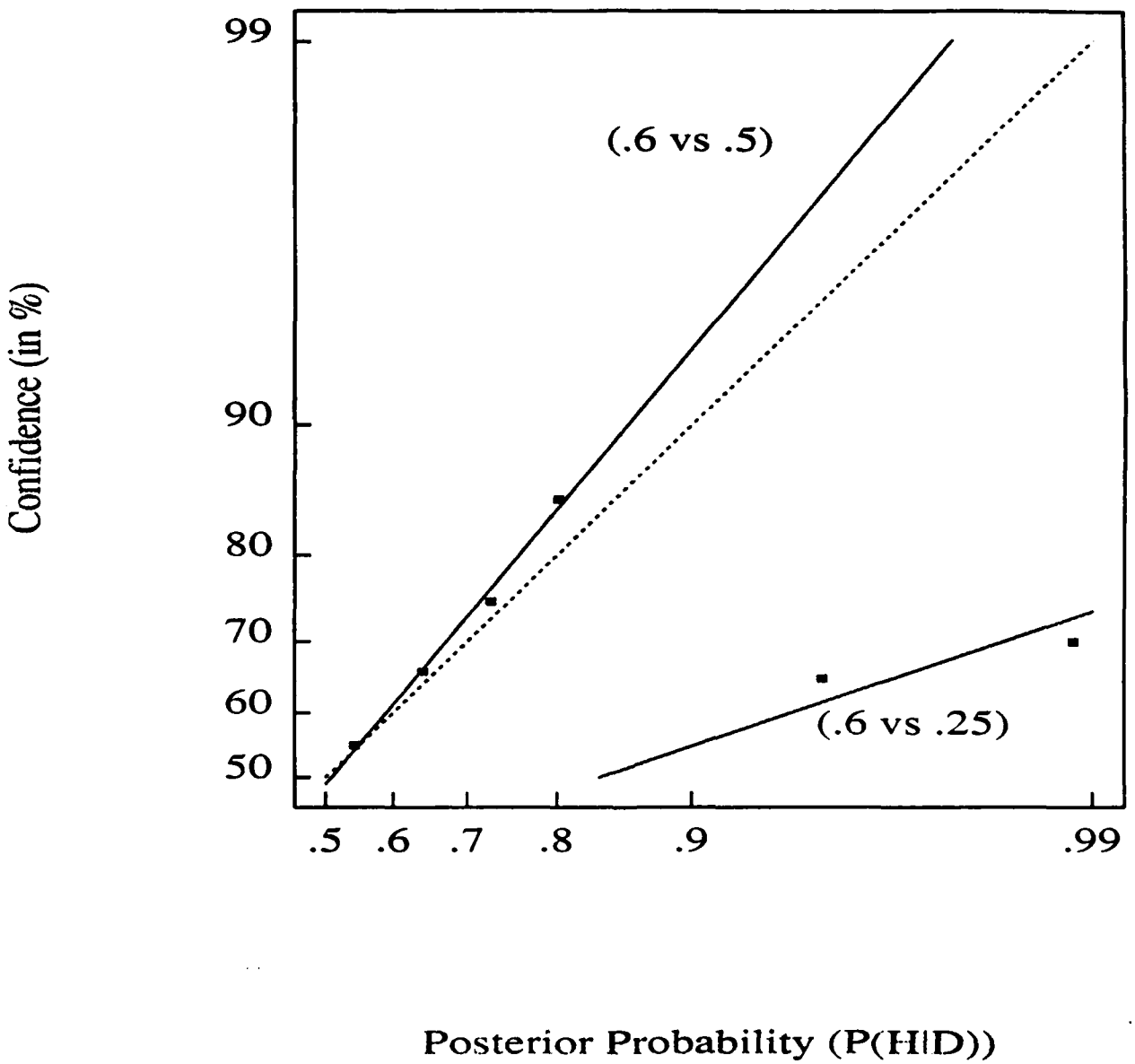
Number of Heads (out of 12)	Separation of Hypotheses (d')	Posterior Probability P(H|D)	Median Confidence (in %)
7	.6 vs .5	.54	55.0
8	.6 vs .5	.64	66.0
9	.6 vs .5	.72	75.0
10	.6 vs .5	.80	85.0
7	.6 vs .25	.95	65.0
8	.6 vs .25	.99	70.0
9	.6 vs .25	.998	80.0
10	.6 vs .25	.999	90.0

hypothesis. We do not plot equal support lines for discriminability and strength, because in this case there is no simple decomposition of posterior probability. Figure 9 plots subjects' median confidence judgments against the Bayesian posterior probability both for low discriminability and high discriminability comparisons. When the discriminability between the hypotheses was low (when the coin's bias was either .6 or .5) subjects were slightly overconfident, when the discriminability between the hypotheses was high (when the bias was either .6 or .25) subjects were grossly underconfident.

Insert Figure 9 about here

The preceding analysis may shed some light on an apparent inconsistency in the hypothesis-testing literature (Fischhoff & Beyth-Marom, 1983). On the one hand, there is considerable evidence that intuitive judgment is subject to a "confirmation bias" (Snyder & Swann, 1978; Wason, 1960) leading to overconfidence in the focal hypothesis. A confirmatory search strategy generates evidence in favor of the focal hypothesis, which produces overconfidence if the fit of the data to the alternate hypothesis is neglected. People are relatively confident that a target is an extrovert after asking what the target would do "to liven up a party" (Snyder & Swann, 1978) presumably because they are not aware that any reasonable answer to this question is consistent with the extroversion hypothesis. Similarly, people make overly confident trait attributions in situations in which an alternate situational explanation can completely account for the data. For example, people attribute a pro-Castro attitude to a person who wrote a pro-Castro essay even though all subjects who were instructed to write such an essay did it regardless of

Figure 9


99
90
80
70
60
50
Confidence (in %)
(.6 vs .5)
(.6 vs .25)
.5
.6
.7
.8
.9
.99
Posterior Probability (P(H|D))

their political position (Jones & Harris, 1967; Ross, 1977). Non-diagnostic information, therefore, increases people's confidence in the focal hypothesis.

On the other hand, there are situations in which the addition of non-diagnostic information reduces people's confidence in the focal hypothesis. For example, Troutman & Shanteau (1977) asked subjects to determine, on the basis of a sample, whether a box of beads contained a majority of white beads (W). They were informed that the box contained either 70 white, 30 red, and 50 blue beads, or 70 red, 30 white, and 50 blue beads. After subjects drew two blue beads, their confidence in the focal hypothesis (W) declined, presumably because this outcome is not representative of W, even though the outcome was non-diagnostic because it was equally probable under the two hypotheses. A similar result in the realm of social judgment has been described by Nisbett, Zukier and Lemley (1981) as the "dilution effect". People's confidence that a target person was a child abuser was reduced when a non-diagnostic item (e.g., he manages a hardware store) was added to his personality profile.

We suggest that both the confirmation bias (e.g., inferring that someone is an extrovert after asking questions likely to yield evidence in favor of extroversion regardless of the target's personality) and the dilution effect (e.g., becoming less confident that a person is a child abuser after learning that the person manages a hardware store) are manifestations of the same psychological phenomenon. People attend primarily to the degree to which the evidence fits the focal hypothesis with insufficient regard for the compatibility between the evidence and the alternate hypothesis. This can produce both overconfidence for information of low or moderate diagnosticity and underconfidence for information of high diagnosticity. For example, Lichtenstein & Feeney (1968) asked subjects to estimate the probability that a dropped "bomb" had been aimed

at one of two cities. When a bomb dropped near one city, subjects were slightly overconfident that the nearby city was the target. However, when a bomb fell far from one city and even farther from the other city, subjects were dramatically underconfident because they did not realize that while the likelihood of the data given City A was quite small, the likelihood of the data given City B was considerably smaller. Similarly, Slowiaczek, Klayman, Sherman & Skov (1989) asked subjects to assess the likelihood that a creature with certain characteristics was a member of a particular species. As in other studies (e.g., Trope & Mackie, 1983), subjects preferred to ask highly diagnostic questions, but they were not sufficiently sensitive to the diagnosticity of the answers. When an answer was highly diagnostic (much more likely under the focal hypothesis than the alternate hypothesis) people were underconfident; when an answer was only weakly or moderately diagnostic (only slightly more likely under one hypothesis than the other) people were overconfident.

The Determinants of Confidence

Study 1 showed that people are generally overconfident in their predictions of future events and underconfident in predicting their own responses. Our analysis of the intuitive weighting of evidence suggests an explanation for this pattern. We have shown that people are much more sensitive to the strength or extremeness of evidence than to its weight or credence. Consequently, overconfidence occurs when strength is high and weight is low, and underconfidence occurs when weight is high and strength is low. Unlike Study 2, where strength and weight were defined by sample proportion and sample size, respectively, and manipulated experimentally, Study 1 does not permit the simple decomposition of evidence into separable components of strength and weight. Nevertheless, the strength-weight distinction may be

invoked to explain the results of evidential problems such as those used in Study 1.

When people assess the probability of some future event (e.g., that the Forty-Niners will win their next two games) they evaluate, we suggest, their impression of the relevant evidence (e.g., the relative strength of the Forty-Niners and their opponents), which may be more or less solid depending on their knowledge. In an analogy to a chance set-up, the extremeness of the impression corresponds to sample proportion, whereas the credence of the impression corresponds to sample size. As in a chance set-up, one could form a strong impression of low credence (e.g., a strong recommendation based on a brief interview), as well as a moderate impression of high credence. Previous studies of the psychology of prediction have shown that intuitive judgments are based primarily on the strength of impression with little or no weight for its credence. In particular, Kahneman and Tversky (1973) demonstrated that people make "predictions by evaluation". These authors presented subjects with paragraph-length descriptions of college freshmen allegedly written by a college counselor on the basis of an interview administered to the entering class. An *evaluation* group was asked to estimate the percentage of students in the class with more favorable descriptions, and a *prediction* group was asked to predict the percentage of students in the class who would obtain a higher freshman GPA. The responses of the two groups were essentially identical although the former group merely evaluated the extremeness of their impression whereas the latter group predicted a remote objective criterion on the basis of limited evidence. Normatively, the prediction group should be considerably more regressive than the evaluation group.

Prediction by evaluation leads to overconfidence whenever people form an extreme impression in a situation where much is unknown (i.e., when an extreme sample proportion is computed from a small sample). The overconfidence observed in the first part of Study 1, as well as in other studies of prediction and general knowledge is consistent with the notion that in these problems, people form strong impressions on the basis of limited evidence. Prediction by evaluation leads to underconfidence when people form a moderate impression on the basis of an extensive body of evidence. In the case of a job choice, for example, underconfidence arises when a person can easily imagine taking either one of the jobs, but fails to appreciate that even a small but clear advantage for Job A over B would lead to the choice of A almost every time. When the balance of arguments in favor of Job A over B is not extreme (say 2 to 1), people's confidence tends to match the balance of arguments, yielding a confidence level of about 2/3. Unless there is a great deal of change in one's state of mind, however, a balance of arguments of 2 to 1 would lead to the choice of A much more often than two-thirds of the time. The tendency to substitute sample proportion, or the relative strength of evidence, for posterior probability has been observed not only in evidential problems but in chance problems as well (Shanteau, 1970). This treatment does not imply that all self predictions will exhibit underconfidence. When asked to predict their own behavior in a novel or unfamiliar situation, for example, people develop strong impressions and are quite overconfident (Vallone, Griffin, Lin, & Ross, in press). Thus we suggest that the pattern of results observed in Study 1 can be explained by the characteristic dominance of strength over weight, and the fact that the weight of the evidence was much greater in the prediction of one's choices than in the prediction of future events. This account is tested in the following study.

Study 6: Self versus other

Twenty-five same sex students, who did not know each other, were given 5 minutes to interview each other. They were told that their task would be to predict each other's behavior in a risky setting. After subjects interviewed each other, they sat at individual computer terminals, where they made predictions and then played a version of the Prisoner's Dilemma game called "The Corporate Jungle". On each move, participants had the option of "merging" their company with their partner's company, or "taking over" their partner's company. If one partner tried to merge and the other tried to take over, the cooperative merger took a steep loss and the corporate raider received a considerable gain. There were 20 payoff matrices, some promoting cooperative behavior and some promoting competitive behavior.

Prior to playing the game, subjects were presented with the 20 payoff matrices. They predicted their own behavior for 10 of the matrices and the behavior of the person they had interviewed for the other 10. They also expressed their confidence (from 50 to 100%) for each prediction. Following the prediction phase, subjects played 20 trials against their opponents, without feedback, and received payment according to their payoffs over the 20 trials.

Subjects were equally confident in their self predictions (M = 84%) and their predictions of others (M = 83%), but they were considerably more accurate in predicting their own behavior (M=81%) than in predicting the behavior of others (M = 68%). In accord with our analysis, people exhibited considerable overconfidence in predictions of others, but were relatively well-calibrated in predicting themselves (see Figure 10).

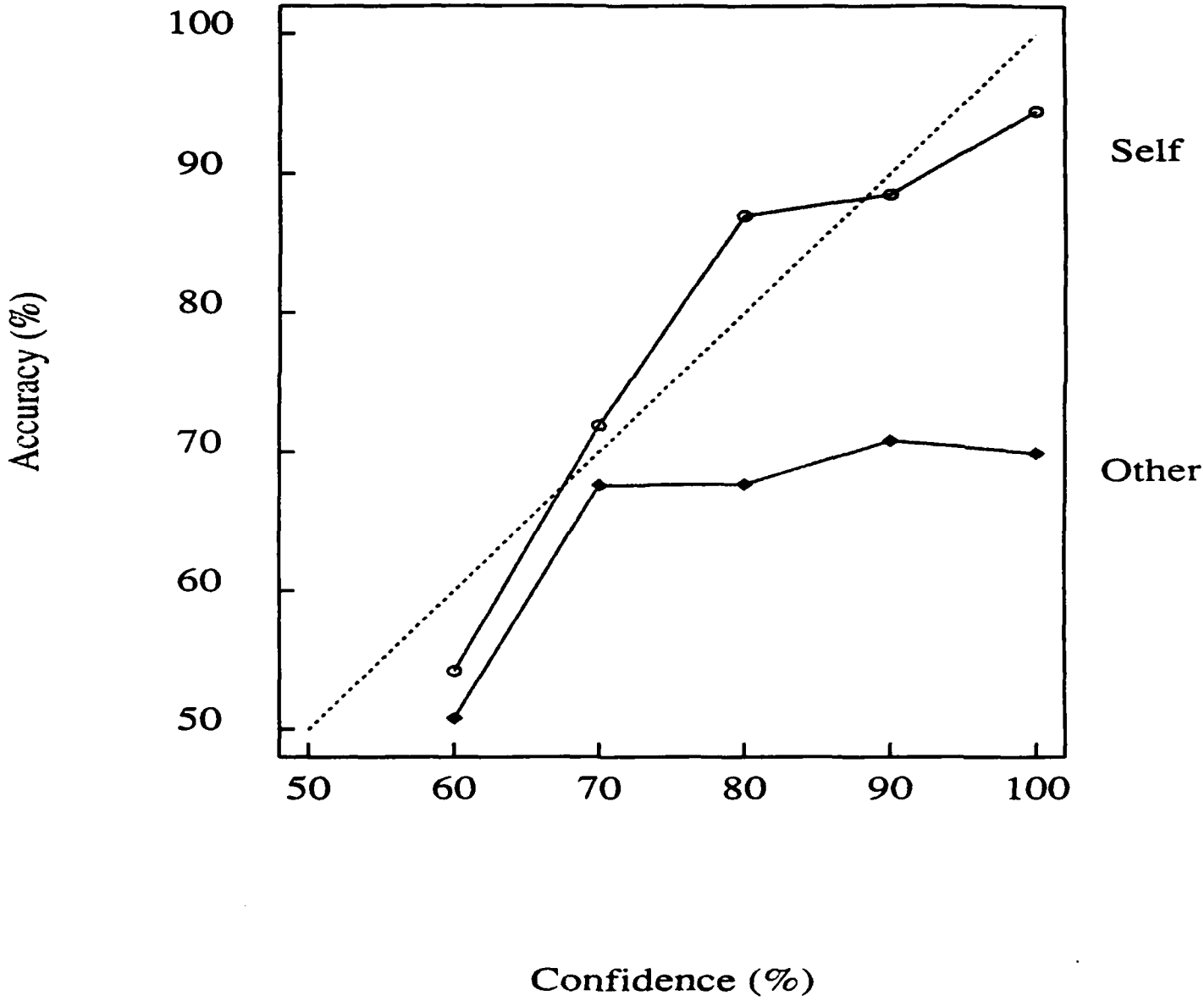
Insert Figure 10 about here

The tendency to be confident about the prediction of the behavior of others, but not of one's own behavior, has been noted by some decision analysts. In decision analysis, one distinguishes decision variables that are controlled by the decision maker and state variables that are not under his or her control. The analysis of decision proceeds by determining the value of decision variables (i.e., decide what you want) and assigning probability to state variables (e.g., the behavior of others). Some decision analysts have noted that their clients wish to follow an opposite course: determine or predict (with certainty) the behavior of others and assign probabilities to their own choices. After all, the behavior of others should be predictable from their traits, needs and interests, whereas our own behavior is highly flexible and contingent on changing circumstances (Jones & Nisbett, 1971).

Concluding Remark

This article explains the observed pattern of overconfidence and underconfidence in terms of dominance relations between evidential variables. This account resolves some apparent inconsistencies concerning the effects of sample size and diagnosticity, and it complements the process-oriented treatment based on judgmental heuristics.

Figure 10


100
90
80
70
60
50
Accuracy (%)
50
60
70
80
90
100
Confidence (%)
Self
Other

The significance of overconfidence to the conduct of human affairs can hardly be overstated. Although overconfidence is not universal, it is prevalent, often massive and difficult to eliminate. This phenomenon is significant not only because it demonstrates the discrepancy between intuitive judgments and the laws of chance, but primarily because confidence controls action (Tversky & Heath, 1989). It has been argued (see e.g., Taylor & Brown, 1988) that overconfidence is adaptive because it moves people to do things that they wouldn't have done otherwise. The advantages of overconfidence, however, may be purchased at a high price. Overconfidence in the diagnosis of a patient, the outcome of a trial, or the projected interest rate could lead to inappropriate medical treatment, bad legal advice and regrettable financial investments. It can be argued that people's willingness to engage in military, legal and other battles would be reduced if they had a more realistic assessment of their chances of success. We doubt that the benefits of overconfidence outweigh its costs.

References

Dawes, R. (1988). *Rational choice in an uncertain world.* New York: Harcourt, Brace, Jovanovich.

Dunning, D., Milojkovic, J., Griffin, D. W., & Ross. L. (in press). The overconfidence effect in social prediction. *Journal of Personality and Social Psychology.*

Edwards, W. (1968). Conservatism in human information processing. In B. Kleinmuntz (Ed.), *Formal representation of human judgment* (pp. 17-52). New York: Wiley.

Fischhoff, B., & Bar-Hillel, M. (1984). Focusing Techniques: A shortcut to improving probability judgments? *Organizational Behavior and Human Performance, 34,* 175-194.

Fischhoff, B., & Beyth-Marom, R. (1983). Hypothesis evaluation from a Bayesian perspective. *Psychological Review, 90,* 239-260.

Fiske, S. T. & Taylor, S. E. (1984). *Social Cognition.* Reading, Mass.: Addison-Wesley.

Forer, B. (1949). The fallacy of personal validation: A classroom demonstration of gullibility. *Journal of Abnormal and Social Psychology, 44,* 118-123.

Grether, D. M. (1980). Bayes rule as a descriptive model: The representativeness heuristic. *The Quarterly Journal of Economics, 95,* 537-557.

Jones, E. E., & Harris, V. A. (1965). The attribution of attitudes. *Journal of Experimental Social Psychology, 3,* 2-24.

Jones, E. E., & Nisbett R. E. (1972). *The actor and the observer: Divergent perceptions of the causes of behavior*. Morristown, N. J.: General Learning Press.

Kahneman, D., Slovic, P., & Tversky, A. (1982). *Judgment under uncertainty: Heuristics and biases*. Cambridge: Cambridge University Press.

Kidd, J. B. (1970). The utilization of subjective probabilities in production planning. *Acta Psychologica, 34*, 338-347.

Kleinbolting, H., & Gigerenzer, G. (1989). *Confidence in one's knowledge: Two kinds?* Manuscript submitted for publication.

Kunda, Z., & Nisbett, R. E. (1986). Prediction and the partial understanding of the law of large numbers. *Journal of Experimental Social Psychology, 22*, 339-354.

Lichtenstein, S., & Feeney, G. F., (1968). The importance of the data-generating model in probability estimation. *Organizational Behavior and Human Performance, 3*, 62-67.

Lichtenstein, S., Fischhoff, B., & Phillips, L. D. (1982). Calibration of probabilities: The state of the art to 1980. In D. Kahneman, P. Slovic, & A. Tversky, (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 306-334). Cambridge: Cambridge University Press.

Lusted, L. B. (1977). *A study of the efficacy of diagnostic radiologic procedures: Final report on diagnostic efficacy*. Chicago: Efficacy Study Committee of the American College of Radiology.

Malsch, M. (1988). *Can lawyers predict the outcome of their cases?* Unpublished manuscript.

May, R. S. (1988). Overconfidence in overconfidence. In A. Chikan, j. Kindler & I. Kiss (Eds.), *Proceedings of the Fourth FUR Conference*. Dordrecht: Kluwer Academic Publishers.

Murphy, A. H., & Winkler, R. L. (1977). Can weather forecasters formulate reliable probability forecasts of precipitation and temperature? *National Weather Digest, 2,* 2-9.

Nisbett, R. E., Zukier, H., & Lemley, R. E. (1981). The dilution effect: Nondiagnostic information weakens the implications of diagnostic information. *Cognitive Psychology, 13,* 248-277.

Oskamp, S. (1965). Overconfidence in case-study judgments. *The Journal of Consulting Psychology, 29,* 261-265.

Ross, L. (1977). The intuitive psychologist and his shortcomings: Distortions in the attribution process. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 10, pp. 174-177). New York: Academic Press.

Shanteau, J. (1970). An additive model for sequential decision making. *Journal of Experimental Psychology, 85,* 181-191.

Slovic, P., Lichtenstein, S., & Fischhoff, B. (1989). Decision Making. In R. C. Atkinson, R. J. Herrnstein, G. Lindzey, & R. D. Luce (Eds.), *Steven's handbook of experimental psychology* (2nd ed.). New York: Wiley.

Slowiaczek, L. M., Klayman, J., Sherman, S. J., & Skov, R. B. (1989). *Information selection and use in hypothesis testing: What is a good question and what is a good answer?* Manuscript submitted for publication.

Snyder, M., & Swann, W. B., Jr. (1978). Behavioral confirmation in social interaction: From social perception to social reality. *Journal of Experimental Social Psychology, 14,* 637-644.

Stael von Holstein, C.-A. S. (1972). Probabilistic forecasting: An experiment related to the stock market. *Organizational Behavior and Human Performance, 8,* 139-158.

Taylor, S. E., & Brown, J. D. (1988). Illusion and well-being: A social psychological perspective on mental health. *Psychological Bulletin, 103,* 193-210.

Trope, Y., & Mackie, D. M. (1987). Sensitivity to alternatives in social hypothesis-testing. *Journal of Experimental Social Psychology, 23,* 445-459.

Troutman, C. M., & Shanteau, J. (1977). Inferences based on nondiagnostic information. *Organizational Behavior and Human Performance, 19,* 43-55.

Tversky, A., & Heath (1989). Ambiguity and confidence in choice under uncertainty. *Proceedings of the 12th research conference on subjective probability, utility, and decision making.* Moscow, USSR.

Tversky, A., & Kahneman, D. (1971). The belief in the "law of small numbers. *Psychological Bulletin, 76,* 105-110.

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology, 5,* 207-232.

Tversky, A., & Kahneman, D. (1980). Causal schemas in judgments under uncertainty. In M. Fishbein (Ed.), *Progress in social psychology.* Hillsdale, N.J.: Erlbaum.

Vallone, R., Griffin, D.W., Lin, S., & Ross, L. (in press). The overconfidence effect in predictions about the self and others. *Journal of Personality and Social Psychology.*

von Winterfeldt, D. & Edwards, W. (1986). *Decision Analysis and Behavioral Research.* New York: Cambridge University Press.

Wason, P. C. (1960). On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology, 12,* 129-140.

Figure Captions

Figure 1. Calibration curve for the prediction of future events.

Figure 2. Calibration curve for the prediction of subjects' own responses.

Figure 3. Equal support lines for strength and weight.

Bayes Rule (dotted lines)

Judged Confidence (solid lines)

Figure 4. Median confidence judgments as a function of posterior probability for a small (n=5) and a large (n=17) sample.

Figure 5. The effect of predictor reliability and criterion reliability on judged confidence (from Kunda and Nisbett, 1986).

Figure 6. Equal support lines for strength and base rate.

Bayes Rule (dotted lines)

Judged Confidence (solid lines)

Figure 7. Median confidence judgments as a function of posterior probability for high (9 to 1) and low (1 to 9) base rates.

Figure 8. Confidence and accuracy in social predictions for high and low base rates (from Dunning et. al, in press).

Figure 9. Median confidence judgments as a function of posterior probability for low discriminability (.6 vs .5) and high discriminability (.6 vs .25).

Figure 10. Calibration curve for the prediction of self and others in a Prisoner's Dilemma game.

Ambiguity and Competence in Choice under Uncertainty

Chip Heath and Amos Tversky
Stanford University

This work was supported by Grant 89-0064 from The Air Force Office of Scientific Research to Stanford University.

Abstract

We investigate the relation between judgments of belief and preferences between bets. It appears that people's willingness to bet on their beliefs depends not only on the perceived likelihood of the event in question (subjective probability) and the precision with which it is measured (ambiguity or vagueness); it also depends on people's general knowledge or competence regarding the relevant domain. We hypothesize that, holding belief or judged probability constant, people prefer to bet in areas where they feel knowledgeable or competent and they avoid betting in domains where they feel ignorant or uninformed. We assume that the feeling of competence is enhanced by knowledge and experience, and it is reduced by emphasizing relevant information that is not available to the decision maker, especially if it is available to others. This account, called the competence hypothesis, can explain the preference to bet on clear over vague probabilities in a chance setup as well as other observed phenomena that cannot be explained by ambiguity, such as the preference to bet on the future over the past, and the preference for skill over chance.

The competence hypothesis is supported in a series of experiments showing that people prefer to bet on their belief over a matched chance event when they feel knowledgeable or competent, and they prefer to bet on chance events over their judgment when they feel ignorant or uninformed. Moreover, they are paying in effect a 20% premium to bet on the more familiar propositions. These results cannot be explained by reference to ambiguity or second-order probability. The implications of the present findings to choice under uncertainty and the measurement of belief are discussed in the last section.

The uncertainty we encounter in the world is not readily quantified. We may feel that our favorite football team has a good chance to win the championship match, that the price of gold will probably go up, and that the incumbent mayor is unlikely to be re-elected, but we are normally reluctant to assign numerical probabilities to these events. However, to facilitate communication and enhance the analysis of choice we are sometimes asked to express our beliefs in numerical form. This task requires a mapping of an impression or a mental state into the language of chance. When we say that the chance of an uncertain event is 30%, for example, we express the belief that we consider this event to be as probable as the drawing of a red ball from a box that contains 30 red and 70 green balls. Does this thought experiment provide an adequate method for measuring belief or subjective probability? What tests can be performed to ensure the meaningfulness of these numbers?

Aside from reliability and internal consistency, proper subjective probabilities must satisfy an additional assumption, which may be called *source independence*. This condition says that if the judged probability of an uncertain event E is P, then the decision maker should be as willing to bet $X on the occurrence of E or on the drawing of a red ball from a box in which the proportion of red balls is P. Furthermore, if the decision maker regards two propositions (e.g., regarding sport and politics) as equally likely, he or she should be equally willing to bet on either one. Thus, preferences between risky prospects depend on the degree of uncertainty but not on its source. (This need not hold for events whose occurrence is desirable or undesirable in itself, such as a win or a loss of one's favorite football team.) The assumption of source independence is implicit in the work of Ramsey (1931) and Savage (1954) that provides the foundation for the modern theory of utility and subjective probability. Indeed, it is the basis for using preference

between bets to measure belief, or subjective probability.

Subjective expected utility theory and the assumption of source independence were challenged by Daniel Ellsberg (1961; see also Fellner, 1961) who constructed a compelling demonstration of what has come to be called an ambiguity effect, although the term "vagueness" may be more appropriate. The simplest demonstration of this effect involves two boxes: one contains 50 red balls and 50 green balls, whereas the second contains 100 red and green balls in unknown proportion. You draw a ball blindly from a box and guess its color. If your guess in correct, you win $20, otherwise you get nothing. On which box would you rather bet? Ellsberg argued that people prefer to bet on the 50/50 box than on the box with the unknown composition, even though they have no color preferences so they are indifferent between betting on red or on green in either box. This pattern of preferences, which has been later demonstrated in many experiments, violates the additivity of subjective probability because it implies that the sum of the probabilities of red and of green is higher in the 50/50 box than in the unknown box.

Ellsberg's work has generated a great deal of interest for two reasons. First, it provides a compelling counter-example to (subjective) expected utility theory within the context of games of chance. Second, it suggests a general hypothesis that people prefer to bet on clear than on vague events, at least for moderate and high probability. For small probability, Ellsberg suggested, people may prefer vagueness to clarity. These observations present a serious problem for expected utility theory and other models of risky choice because, with the notable exception of games of chance, most decisions in the real world depend on uncertain events whose probabilities cannot be precisely assessed. This is especially true for probabilities based on intuitive judgment that are generally approximate and vague. If people's choices depend not only on the

degree of uncertainty but also on the precision with which it can be assessed, then the applicability of the standard models of risky choice is severely limited. Indeed, several authors have extended the standard theory by invoking nonadditive measures of belief, and second-order probability distributions in order to account for the effect of ambiguity. The normative status of these models is a subject of a lively debate. Several authors, notably Ellsberg (1963), maintain that aversion to ambiguity can be justified on normative grounds, although Raiffa (1961) has shown that it can lead to incoherence.

Ellsberg's example and most of the subsequent experimental research on the response to ambiguity or vagueness has been confined to chance processes, such as drawing a ball from a box, or to situations in which the decision maker is provided with a probability estimate. The potential significance of ambiguity, however, stems from its relevance to the evaluation of evidence in the real world. This observation raises the question of whether ambiguity aversion is confined to games of chance and stated probability, or whether it also holds for judgmental probabilities based on knowledge rather than on considerations of symmetry or total ignorance. We found no answer to this question in the literature, but there is evidence that casts some doubt on the generality of this phenomenon.

For example, Budescu, Weinberg, and Wallsten (1986) compared the cash equivalents of gambles whose probabilities were expressed numerically, graphically, or verbally. In the graphical display, probabilities were presented as the shaded area of a circle. In the verbal form, probabilities were described by expressions such as "very likely" or "highly improbable". Because the verbal and the graphical forms are more ambiguous than the numerical form, ambiguity aversion implies a preference for the numerical over the other displays. This prediction was not

confirmed. Subjects priced the gambles roughly the same in all three displays. In a different experimental paradigm, Cohen and Hansel (1959) and Howell (1971) investigated subjects' choices between compound gambles involving both skill and chance components. For example, in the latter experiment the subject has to hit a target with a dart (on which the subject's hit rate = 75%) as well as spin a roulette wheel so that it will land on a marked section comprising 40% of the area. Success involves a 75% skill component and 40% chance component with an overall probability of winning of $.75 \times .4 = .3$. Howell (1971) varied the skill and chance components of the gambles, holding the overall probability of winning constant. Because the chance level was known to the subject whereas the skill level was not, ambiguity-aversion implies that subjects would shift as much uncertainty as possible to the chance component of the gamble. In contrast, 87% of the choices reflect a preference for skill over chance. Cohen and Hansel (1959) obtained essentially the same result.

The Competence Hypothesis

The preceding discussion suggests that the aversion to ambiguity observed in a chance setup (involving aleatory uncertainty) does not readily extend to judgmental problems (involving epistemic uncertainty). In this article, we investigate an alternative account of uncertainty preferences, called the competence hypothesis, which applies to both chance and evidential problems. We submit that the willingness to bet on an uncertain event depends not only on the estimated likelihood of that event and the precision of that estimate; it also depends on one's general knowledge or understanding of the relevant context. More specifically, we propose that -- holding degree of belief or judged probability constant -- people prefer to bet in a context

where they consider themselves knowledgeable or competent than in a context where they feel ignorant or uninformed. We assume that the feeling of competence in a given context is enhanced by general knowledge, familiarity and experience, and it is diminished, for example, by calling attention to relevant information that is not available to the decision maker, especially if it is available to others.

There are both judgmental and preferential reasons for the competence hypothesis. First, people may have learned from a life-long experience that they generally do better in situations they understand or control than in situations in which they have less knowledge and competence. Thus, they may expect to do better in the former case, and this feeling may carry over to situations where the chances of winning are no longer higher in the familiar than in the unfamiliar context. Second, people may like to bet on their (physical or mental) skills, either because they enjoy the challenge or because they like to demonstrate their competence. Conversely, people may avoid a situation they do not understand because it makes them feel incompetent.

The competence hypothesis readily applies to Ellsberg's example. People do not like to bet on the unknown box, we suggest, because there exists relevant evidence, namely the proportion of red and green balls in the box, that is knowable in principle but unknown to them. The presence of such data make people feel less knowledgeable and less competent and reduce the attractiveness of the corresponding bet. A closely related interpretation of Ellsberg's example has been offered by Frisch and Barron (1988). The present account is also consistent with the finding of Curley, Yates, and Abrams (1986) that the aversion to ambiguity is enhanced by the anticipation that the contents of the unknown box will be shown to others.

Essentially the same analysis applies to the preference for betting on the future rather than on the past. Rothbart and Snyder (1970) asked subjects to roll a die and bet on the outcome either before the die was rolled or after the die was rolled but before the result was revealed. The subjects who predicted the outcome before the die was rolled expressed greater confidence in their guesses than the subjects who postdicted the outcome after the die roll. The prediction group also bet significantly more money than the postdiction group. The authors attributed this phenomenon to magical thinking, the belief that the subjects can exercise some control over the outcome before, but not after, the roll of the die. However, the preference to bet on future rather than chance events is observed even when magical thinking is unlikely, as illustrated by the following problem in which subjects were presented with a choice between the two bets:

a) A stock is selected at random from the Wall Street Journal. You guess whether it will go up or down tomorrow. If you're right, you win $5.

b) A stock is selected at random from the Wall Street Journal. You guess whether it went up or down yesterday. You cannot check the paper. If you're right you win $5.

Sixty-seven percent of the subjects (N=184) preferred to bet on tomorrow's closing price than on yesterday's closing price. (Ten percent of the participants, selected at random, actually played their chosen bet.) Because the past -- unlike the future -- is known to others but not to themselves, subjects prefer to bet on the future where their relative ignorance is lower. Similarly, Brun and Teigen (1989) observed that subjects preferred to guess the result of a die roll, the sex of a child or the outcome of a soccer game before the event rather than afterward. Most of the subjects found guessing before the event more "satisfactory if right" and less "uncomfortable if wrong." In prediction, only the future can prove you wrong; in postdiction, you could be wrong right now. The same argument applies to Ellsberg's problem. In the 50/50 box, a guess

could turn out to be wrong only after drawing the ball. In the unknown box, on the other hand, the guess may turn out to be mistaken even before the drawing of the ball -- if it turns out that the majority of balls in the box are of the opposite color. It is noteworthy that the preference to bet on future rather than on past events cannot be explained in terms of ambiguity because there is no reason to believe that, in these problems, the future is less ambiguous than the past -- it is merely less knowable.

Simple chance events, such as drawing a red ball from a box containing a specified number of red and green balls involves no ambiguity; the chance of winning are known precisely. If betting preferences between equiprobable events are determined primarily by considerations of ambiguity, people should prefer to bet on chance over their own vague judgments, at least for moderate and high proability. The present account predicts a different pattern. Although people feel more knowledgeable facing the 50-50 box than the unknown box, many do not feel especially competent to deal with chance, either because they do not really understand how it operates or because they feel that others may be luckier than they are. Be that as it may, we suggest that people often feel more knowledgeable and more competent in dealing with familiar, uncertain phenomena, such as sport or the weather, despite the fact that the probabilities of such events are vague. The present analysis implies that people will prefer betting on their judgment over a matched chance event when they feel knowledgeable and competent, but not otherwise. This prediction is confirmed by the findings of Cohen and Hansel (1959) and of Howell (1971) that people prefer betting on their skill rather than on chance. It is also consistent with the observation of March and Shapiro (1987) that many top managers express aversion to betting on chance events although they consistently bet on highly uncertain business propositions.

We have argued that the competence hypothesis can account for the available evidence on uncertainty preferences, whether or not they involve ambiguity. These include (i) the preference for betting on the known rather than on the unknown box in Ellsberg's problem, (ii) the preference to bet on future rather than on past events, and (iii) the preference for betting on skill rather than on chance. Contrary to the assumption of source independence, which ensures the consistency of preferences and beliefs, the competence hypothesis implies a choice-judgment discrepancy, namely a preference to bet on A rather than on B even though B is judged to be at least as probable as A. In the following series of experiments we test the competence hypothesis and investigate the choice-judgment discrepancy. The relation between judgments of belief and revealed preferences are discussed in the last section.

Experiment 1: Betting on Knowledge

Subjects answered 30 knowledge questions in two different categories, such as, history, geography or sport. Four alternative answers were presented for each question, and the subject first selected a single answer and then rated his or her confidence in that answer on a scale from 25% (pure guessing) to 100% (absolute certainty). Participants were given detailed instructions about the definition of the scale and the notion of calibration. Specifically, they were instructed to use the scale so that a confidence rating of 60%, say, will correspond to a hit rate of 60%. They were also told that these ratings would be the basis for a money-making game, and warned that both underconfidence and overconfidence would reduce their earnings.

After answering the questions and assessing confidence, subjects were given an opportunity to choose between betting on their answers or on a lottery in which the probability of winning was equal to their stated confidence. For a confidence rating of 75%, for example, the subject was given the choice between (i) betting that his or her answer was correct, or (ii) betting on a 75% lottery, defined by drawing a numbered chip in the range 1-75 from a bag filled with 100 numbered poker chips. For half of the questions, lotteries were directly equated to confidence ratings. For the other half of the questions, subjects chose between the complement of their answer (betting that an answer other than the one they chose is correct) or the complement of their confidence rating. Thus, if a subject chose answer "A" with confidence of 65%, the subject could choose between betting that one of the remaining answers "B", "C", or "D" is correct, or betting on a 100% – 65% = 35% lottery.

Two groups of subjects participated in the experiment. One group (N=29) included psychology students who received course credit for participation. The second group (N=26) was recruited from introductory economic classes and performed the experiment for cash earnings. To determine the subjects' payoffs, ten questions were selected at random, and the subjects played out the bets they had chosen. If subjects chose to gamble on their answer, they collected $1.50 if their answer was correct. If subjects chose to bet on the chance lottery, they drew a chip from the bag and collected $1.50 if the number on the chip fell in the proper range and nothing otherwise. Average earnings for the experiment were around $8.50.

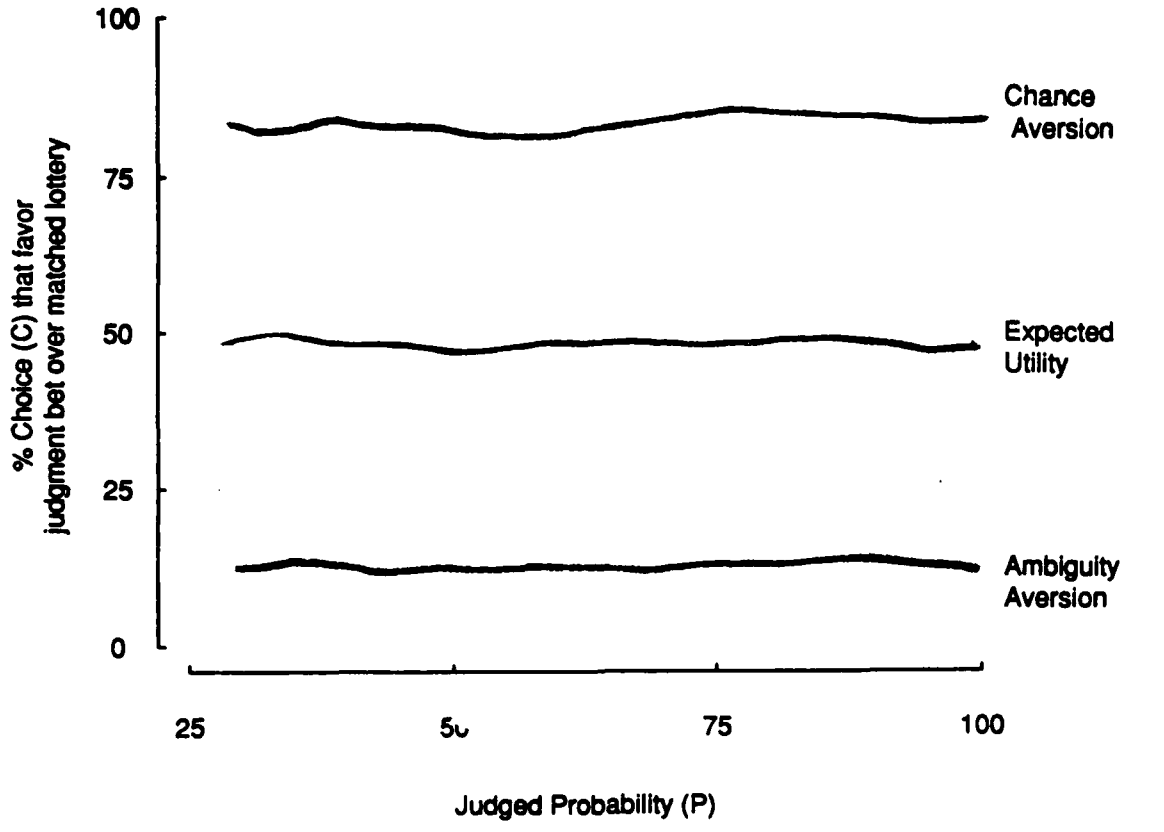
Paid subjects took more time than unpaid subjects in selecting their answers and assessing confidence; they were slightly more accurate. Both groups exhibited overconfidence: the paid subjects answered correctly 47% of the questions and their average confidence was 60%.

The unpaid subjects answered correctly 43% of the questions and their average confidence was 53%. (Only the data from the simple lotteries are reported in subsequent analyses. The complementary lotteries were added primarily to balance the payoffs.)

The results are summarized by plotting the percentage of choice C that favor the judgment bet over the lottery, as a function of judged probability P. Before discussing the actual data, it is instructive to examine several contrasting predictions, implied by five alternative hypotheses, which are displayed in Figure 1.

Insert Figure 1 about here

The upper panel of Figure 1 displays the predictions of three hypotheses in which C is independent of P. According to expected utility theory, decision makers will be indifferent between betting on their judgment or betting on a chance lottery, hence C should equal 50% throughout. Ambiguity aversion implies that people will prefer to bet on a chance event whose probability is well defined than on their judged probability, which is inevitably vague, hence C should fall below 50% everywhere. The complementary hypothesis, called chance aversion, predicts that people will prefer to bet on their judgment than on a matched chance lottery, hence C should exceed 50% for all P. In contrast to the flat predictions displayed in the upper panel, the two hypothesis in the lower panel imply that C depends on P. The regression hypothesis states that the decision weights, which control choice, will be regressive relative to stated probabilities. Thus, C will be relatively high for small probabilities, and relatively low for high proba-



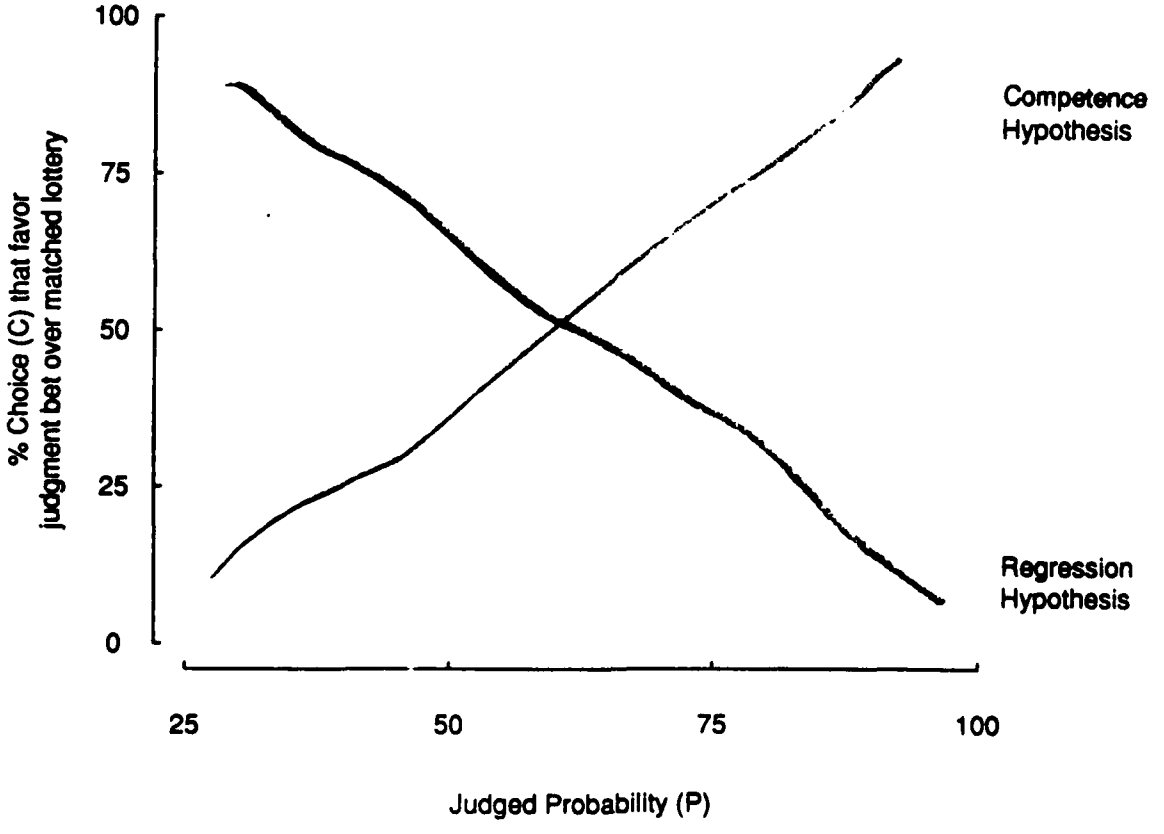


Figure 1. Five contrasting predictions of the results of an uncertainty preference experiment.

bilities. This prediction also follows from the theory put forth by Einhorn and Hogarth (1985). These authors hypothesized a particular process model based on mental simulation, adjustment and anchoring but their predictions coincide with the regression hypothesis. Finally, the competence hypothesis, introduced in this paper, implies that people will tend to bet on their judgment when they feel knowledgeable and on the chance lottery when they feel ignorant. Because higher stated probability generally entails higher knowledge, C will be an increasing function of P, except at 100% where the chance lottery amounts to a sure thing.

Insert Figure 2 about here

Insert Table 1 about here

The results of the experiment are summarized in Table 1 and Figure 2. Table 1 presents, for three different ranges of P, the percentage of paid and non-paid subjects who bet on their answers rather than on the matched lottery. Recall that each question had four possible answers so the lowest confidence level is 25%. Figure 2 displays the overall percentage of choices C that favored the judgment bet over the lottery as a function of judged probability P. (In this and all subsequent figures, we plot the isotone regression of C on P. That is, the best-fitting monotone function in the least squares sense, see Barlow, Bartholomew, Brimmner & Brunk, 1972). The graph shows that the subjects chose the lottery when P was low or moderate (below 65%), and

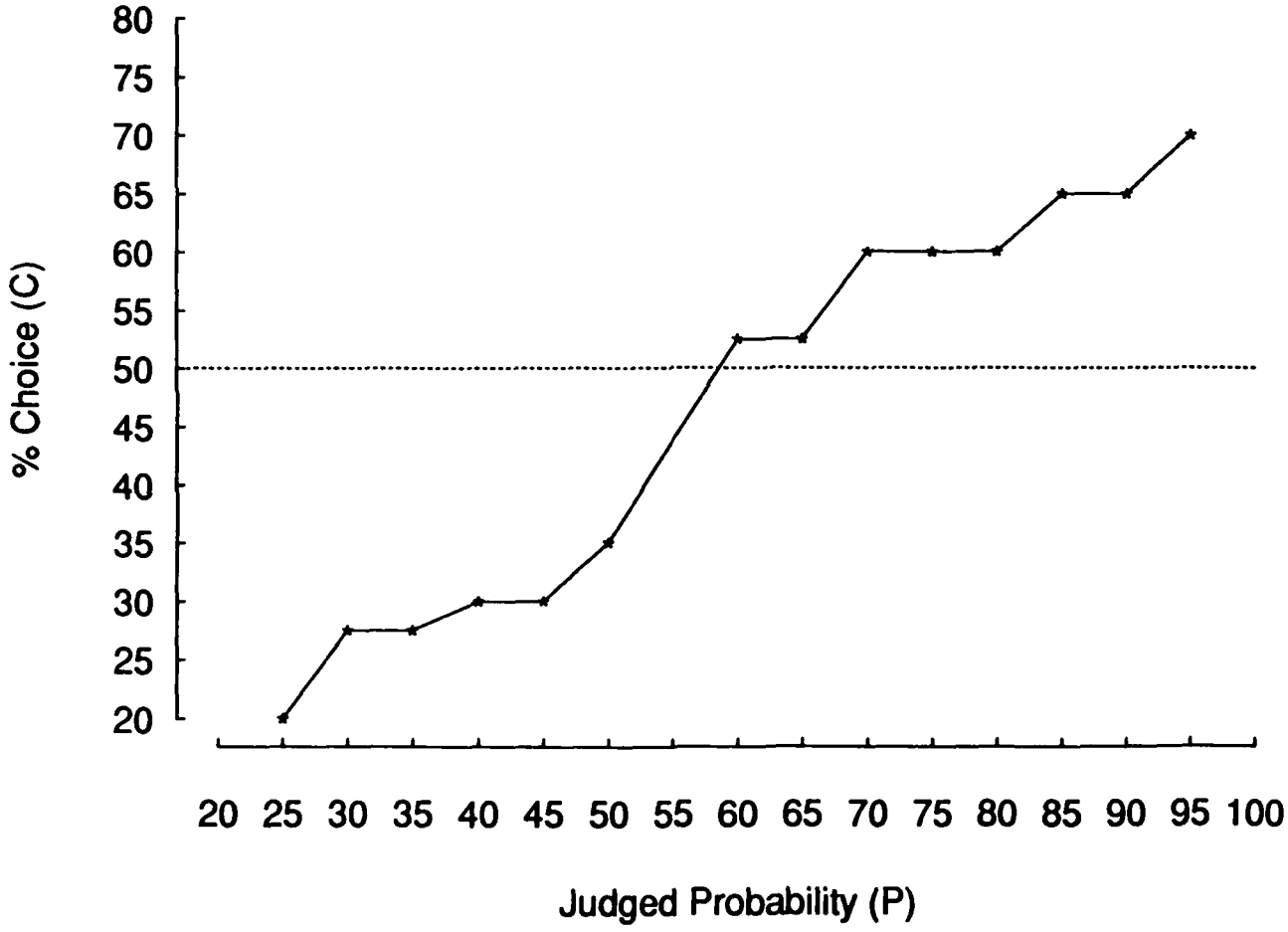


Figure 2. Percentage of choices C that favor a judgment bet over a matched lottery as a function of judged probability P (Experiment 1).

Table 1. Percentage of paid and non-paid subjects who preferred the judgment bet over the lottery for low, medium and high P. The number of observations are given in parenthesis.

	$25 \leq P \leq 50$	$50 < P < 75$	$75 \leq P \leq 100$
Paid	29 (278)	42 (174)	55 (168)
Non-paid	22 (394)	43 (188)	69 (140)

that they chose to bet on their answers when P was high. The pattern of results was the same for the paid as for the non-paid subjects but the effect was slightly stronger for the latter group. These results confirm the prediction of the competence hypothesis and reject the four alternative accounts, notably the ambiguity aversion hypothesis, implied by second-order probability models (e.g., Gärdenfors and Sahlin, 1982), and the regression hypothesis, implied by the model of Einhorn and Hogarth (1985). Alternative interpretations of these data are discussed in the last section.

To obtain a statistical test of the competence hypothesis we computed, separately for each subject, the binary correlation coefficient (ϕ) between choice (judgment bet vs. lottery) and judged probability (above .65 vs. below .65). Seventy-two percent of the subjects yielded positive coefficients and the average ϕ was .30, ($t(54) = 4.3$, $p < .01$). To investigate the robustness of the observed pattern, we replicated the experiment with one major change. Instead of constructing chance lotteries whose probabilities matched the values stated by the subjects, we constructed lotteries in which the probability of winning was either 6% higher or 6% lower than the subjects' judged probability. For high-knowledge questions ($P \geq 75\%$), the majority of responses (70%) favored the judgment bet over the lottery even when the lottery offered a (6%) higher probability of winning. Similarly for low confidence questions ($P \leq 50\%$), the majority of responses (52%) favored the lottery over the judgment bet even when the former offered a lower (6%) probability of winning.

Insert Figure 3 about here

Figure 3 presents the calibration curve for the data of Experiment 1. The figure shows that, on the whole, people are reasonably well-calibrated for low probability, but exhibit substantial overconfidence for high probability. The preference for the judgment bet over the lottery for high probability, therefore, cannot be justified on an actuarial basis.

Experiment 2: Football and Politics

Our next experiment differs from the previous one in several respects. First, it concerns the prediction of real-world future events rather than the assessment of general knowledge. Second, it deals with binary events so that the lowest level of confidence is .5 rather than .25 as in the previous experiment. Finally, in addition to the judgments of probability and the choice between the matched bets, subjects also rated their level of knowledge for each one of the predicted items.

A group of 20 students predicted the outcomes of 14 football games each week for 5 consecutive weeks. For each game, subjects selected the team that they thought would win the game and assessed the probability of their chosen team winning. The subjects also assessed, on a 5-point scale, how knowledgeable they were with respect to each game. Following the rating, subjects were asked whether they preferred to bet on the team they chose or on a matched chance lottery. The results summarized in Figure 4 confirm the previous finding. For both high

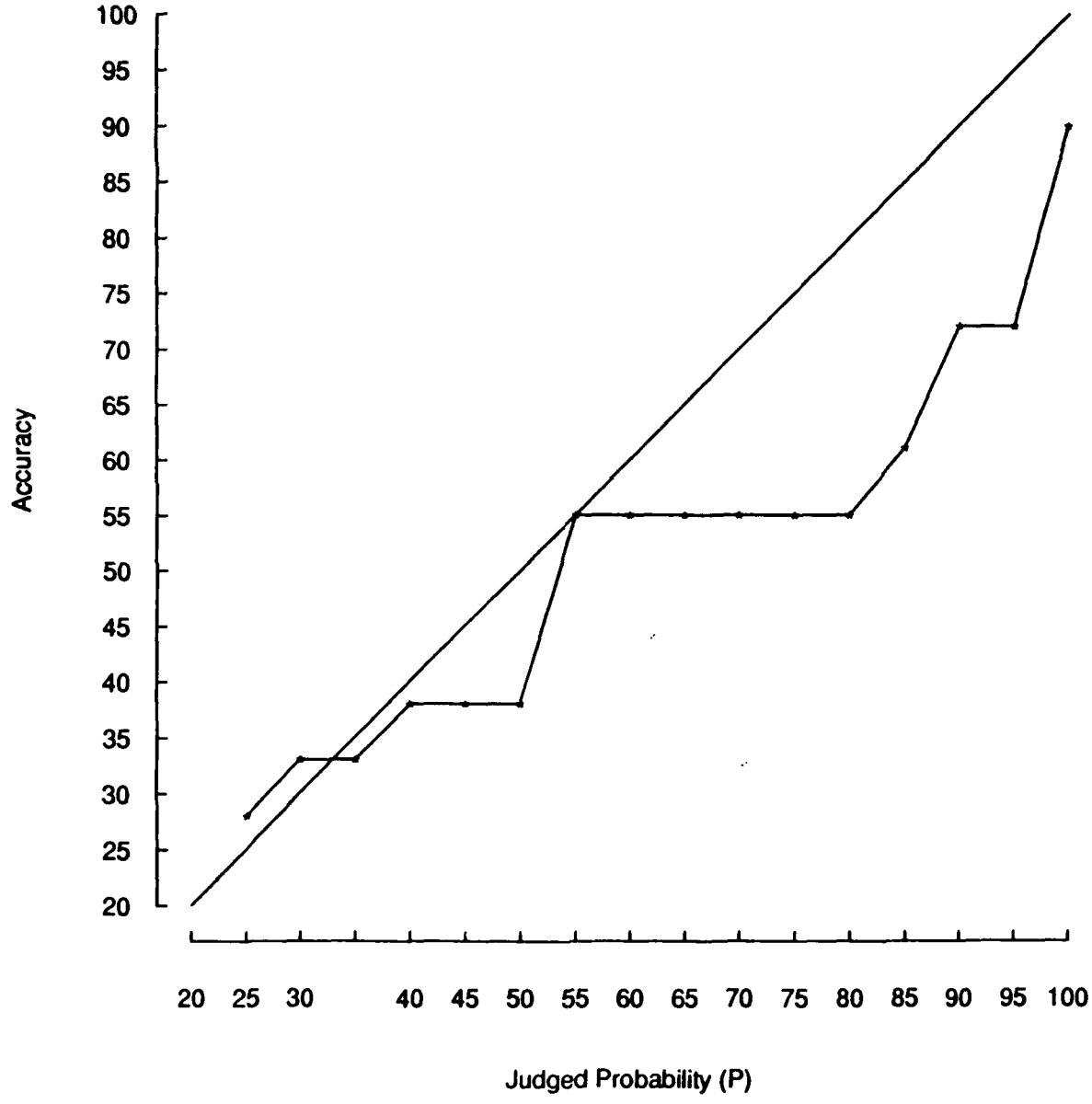


Figure 3. Calibration curve for Experiment 1.

and low knowledge (defined by a median split on the knowledge rating scale), C was an increasing function of P. Moreover, C was greater for high knowledge than for low knowledge at any P > .5. Only 5% of the subjects produced negative correlations between C and P, and the average ϕ coefficient was .33, ($t(77) = 8.7$, $p<.01$).

Insert Figure 4 about here

We next took the competence hypothesis to the floor of the Republican National Convention in New Orleans during August of 1988. The participants were mostly volunteers who worked at the convention. They were given a one-page questionnaire that contained instructions and an answer sheet. Thirteen states were selected to represent a cross-section of different geographical areas as well to include the most important states in terms of electoral votes. The participants (N=100) rated the probability of Bush carrying each of the 13 states in the November 1988 election on a scale from 0 (Bush is certain to lose) to 100 (Bush is certain to win). As in the football experiment, the participants rated their knowledge of each state on a 5-point scale and indicated whether they would rather bet on their prediction or on a chance lottery. The results, summarized in Figure 5, show that C increased with P for both levels of knowledge, and that C was greater for high knowledge than for low knowledge at all levels of P. When asked about their home state, 70% of the participants selected the judgment bet over the lottery. Only 5% of the subjects yielded negative correlations between C and P, and the average ϕ coefficient was .42, ($t(99) = 13.4$, $p<.01$).

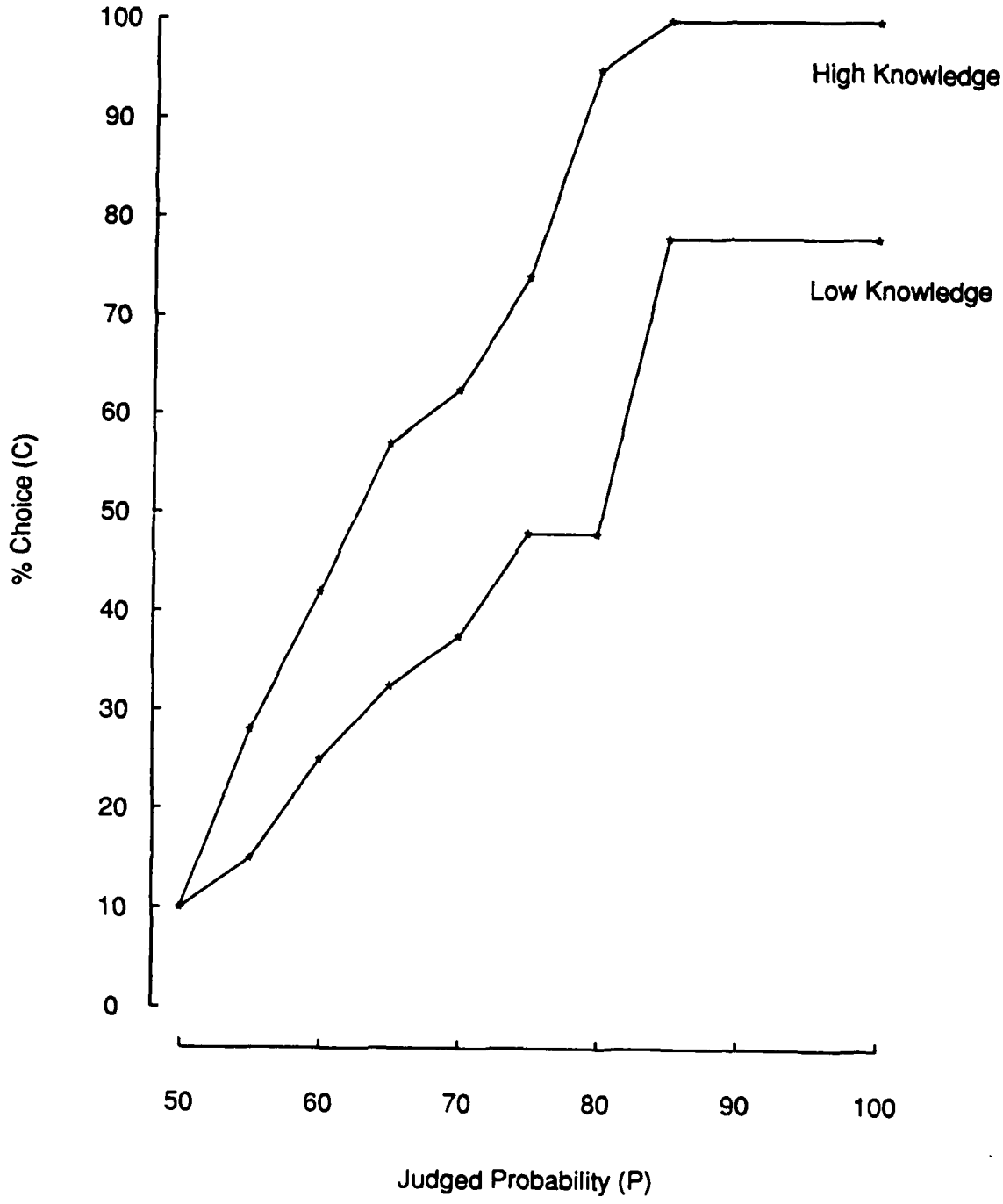


Figure 4. Percentage of choices C that favor a judgment bet over a matched lottery as a function of judged probability P, for high- and low-knowledge items. (Football forecast, Experiment 2).

Insert Figure 5 about here

The results displayed in Figures 4 and 5 support the competence hypothesis in the prediction of real-world events. In both tasks C increases with P as in Experiment 1. Furthermore, the tendency to select the judgment bet over the chance lottery is stronger for the high-knowledge items than for the low-knowledge items throughout the range of judged probability. Recall that in Experiment 1 probability and knowledge (or competence) were perfectly correlated. The choice-judgment discrepancy observed in that experiment, therefore, could be attributed to a distortion of the probability scale in the judgment task. This explanation, however, does not apply to the results of the present experiment, which exhibit a significant effect ($p < .01$ in both experiments) of rated knowledge independent of P. It is noteworthy that the strategy of betting on judgment was less successful than the strategy of betting on chance in both data sets. The former strategy yielded hit rates of 64% and 78% for football and election, respectively, whereas the latter strategy yielded hit rates of 73% and 80%. The observed tendency to select the judgment bet, therefore, does not yield better performance.

Experiment 3: Long Shots

The preceding experiments show that people often prefer to bet on their judgment than on a matched chance event, even though the former is more ambiguous than the latter. This effect summarized in Figures 1, 2 and 3, was observed at the high end of the probability scale.

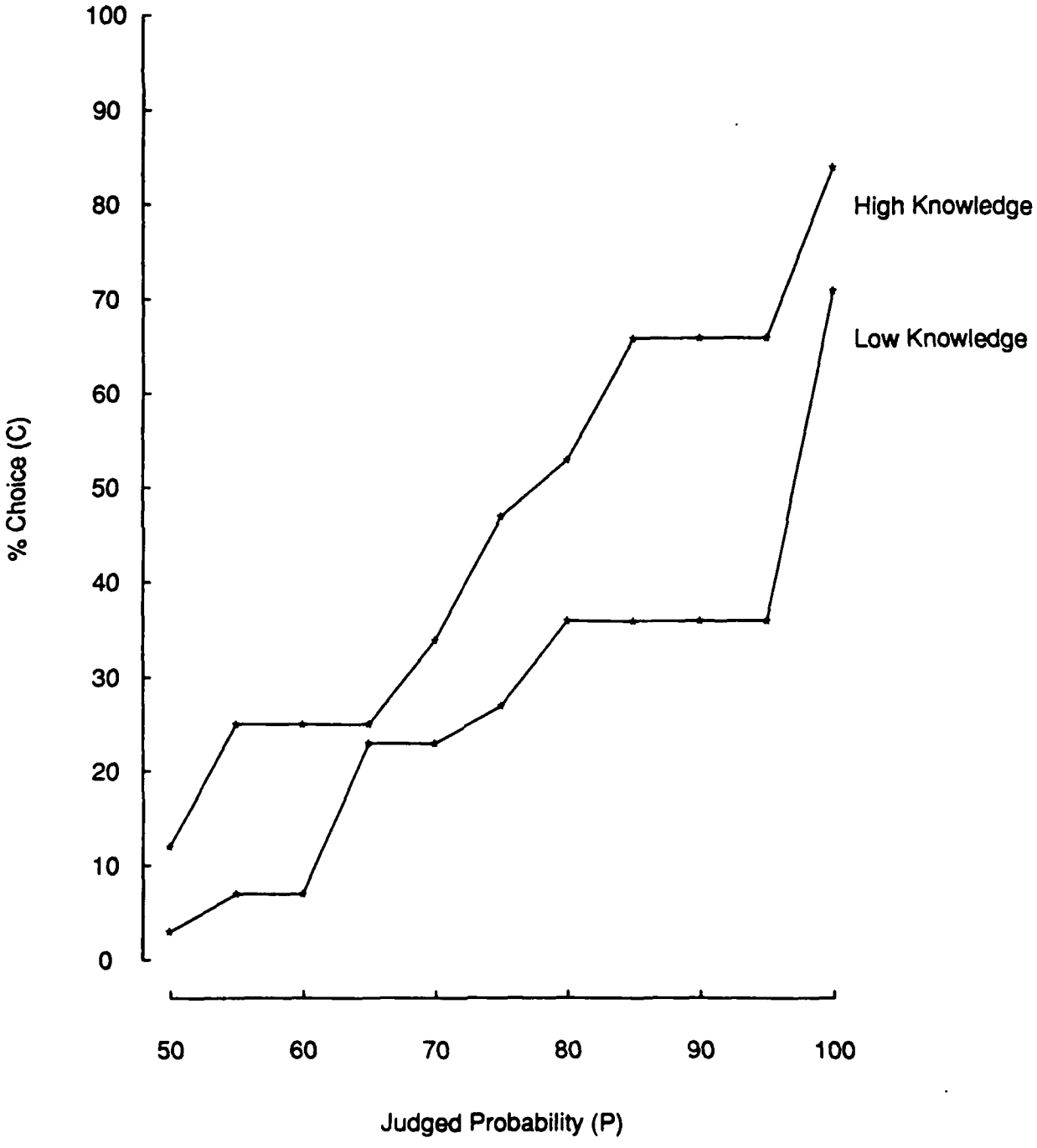


Figure 5. Percentage of choices C that favor a judgment bet over a matched lottery as a function of judged probability P, for high- and low-knowledge items. (Election forecast, Experiment 2).

The present experiment was designed to test whether the preference to bet on one's judgment extends to events whose judged probabilities are relatively low, so the subject's best guess is likely to be wrong.

One hundred and eight Stanford students were presented with open-ended questions about twelve future events (e.g., what movie will win this year's Oscar for best picture? What football team will win the 1990 Super Bowl? In what class next quarter will you have the highest grade?). They were asked to answer each question, to estimate the chances that their guess will turn out to be correct, and to indicate whether they have high or low knowledge of the relevant domain. The use of open-ended questions eliminates the lower bound of 50% imposed by the use of dichotomous predictions in the previous experiment. After the subjects completed these tasks, they were asked to consider, separately for each question, whether they would rather bet on their prediction or on a matched chance lottery.

Insert Table 2 about here

On average, the subjects answered ten out of the twelve questions. Table 2 presents the percentage (C) of responses that favor the judgment bet over the chance lottery for high- and low-knowledge items, and for judged probabilities below or above .5. The number of responses in each cell is given in parentheses. The results show that, for high-knowledge items, the judgment bet was preferred over the chance lottery regardless of whether P was above or below one-half ($p < .01$ in both cases), indicating that for high-knowledge items people prefer to bet on

Table 2. Percentage of choices (C) that favor a judgment bet over a matched lottery for high- and low-rated knowledge and for judged probability below and above .5. The number of responses are given in parentheses.

	Judged Probability	
	$P < .5$	$P \geq .5$
Rated Knowledge:		
Low	36	58
	(593)	(128)
High	61	69
	(151)	(276)

their predictions -- even when they are unlikely to be true. Indeed, the discrepancy between the low- and high-knowledge conditions was greater for $P < .5$ than for $P \geq .5$.

Experiment 4: Expert Prediction

In the preceding experiments, we used the subjects' ratings of specific items to define high and low knowledge. In this experiment, we employ a more general definition of knowledge, or competence, based on the sorting of subjects according to their expertise. To this end, we asked 110 students in an introductory psychology class to assess their knowledge of politics and of football on a 9-point scale. All subjects who rated their knowledge of the two areas on opposite sides of the mid-point were asked to take part in the experiment. Twenty-five subjects met this criterion and all but two agreed to participate. They received course credit for participation and were informed that, on average, they are expected to win an additional $10. The participants included 12 political "experts" and 11 football "experts" defined by their strong area. To induce the subjects to give careful responses, we gave them detailed instructions including a discussion of calibration, and we employed the Brier scoring rule (see e.g., Lichtenstein et. al, 1982) designed to motivate subjects to give their best estimates.

The experiment consisted of two sessions. In the first session, each subject made predictions for a set of 40 future events (20 political events and 20 football games). All the events were resolved within five weeks of the date of the initial session. The political events concerned the winner of the various states in the 1988 presidential election. The 20 football games included 10 professional and 10 college games. For each contest (politics or football), subjects chose a winner by circling the name of one of the contestants, and then assessed the probability

that their prediction would come true (on a scale from 50% to 100%).

Using the results of the first session, 20 triples of bets were constructed for each participant. Each triple included three matched bets with the same probability of winning generated by (i) a chance device, (ii) the subject's prediction in his or her strong category, (iii) the subject's prediction in his or her weak category. In the second session, subjects ranked each of the 20 triples of bets. The chance bets were defined as in Experiment 1 with reference to a box containing 100 numbered chips. Subjects were told that they would actually play their choices in each one of the triples. To encourage careful ranking, subjects were told that they would play 80% of their first choices and 20% of their second choices.

Insert Table 3 about here

Insert Figure 6 about here

The data are summarized in Table 3 and Figure 6, which plots the attractiveness of the three types of bets (inverse mean rank) against judged probability. The results show a strong preference for betting on the strong category. Across all triples, the mean ranks were 1.68 for the strong category, 2.08 for the chance lottery, and 2.23 for the weak category. The difference among the ranks is highly significant ($p<.001$) by the Wilcoxen rank sum test. These results demonstrate that people prefer to bet on their judgment in their area of competence, but prefer to

Table 3. Ranking data for Expert Study.

Rank:	1st	2nd	3rd	Mean Rank
High-Knowledge Bet	192	85	68	1.64
Chance Bet	74	155	116	2.12
Low-Knowledge Bet	79	105	161	2.23

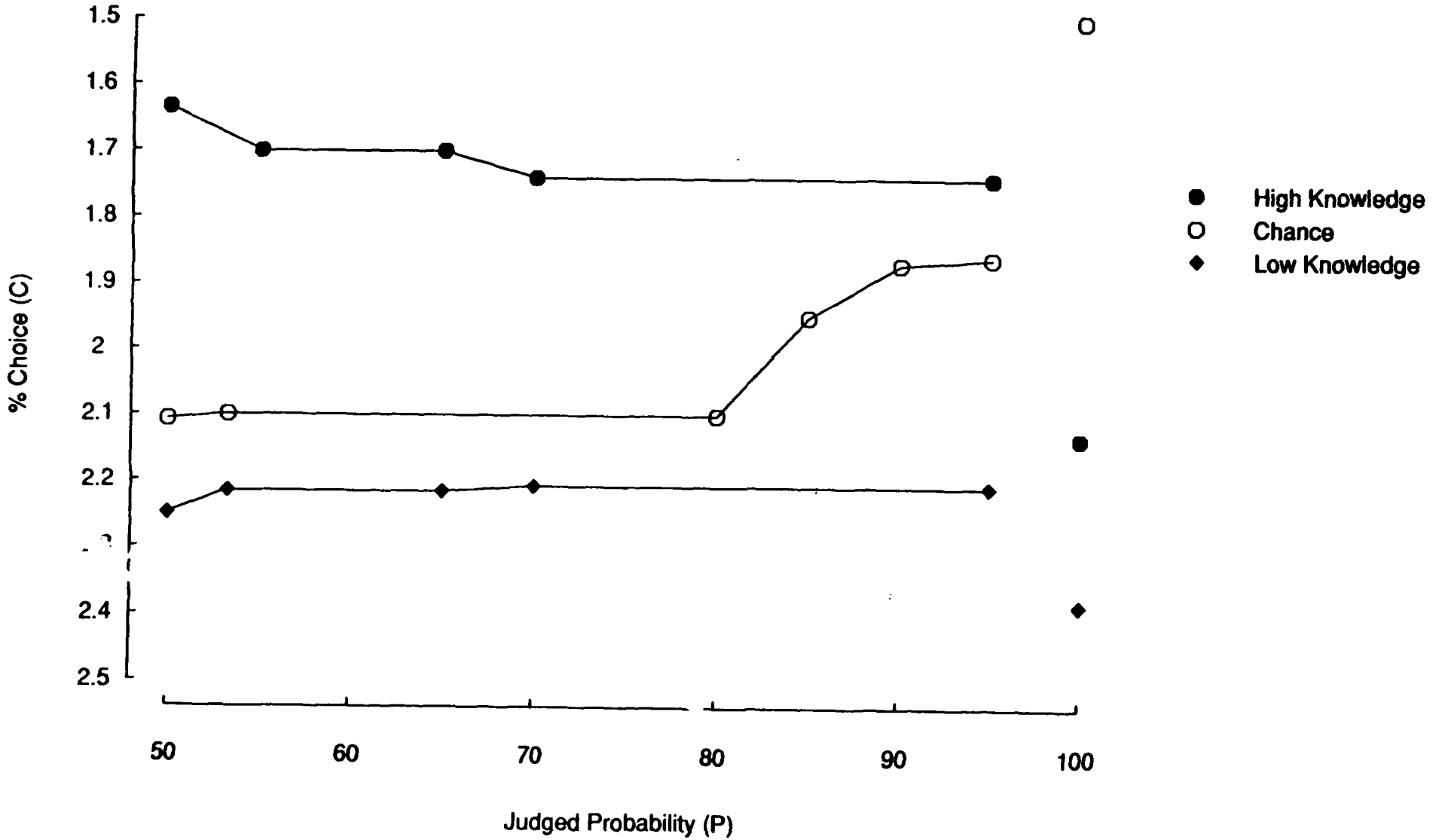


Figure 6. Ranking data for high knowledge, low knowledge and chance bets as a function of P.

bet on chance in an area in which they are no well-informed. As expected, the lottery became more popular than the high-knowledge bet only at 100%. This pattern of result is inconsistent with an account based on ambiguity or second-order probabilities because both the high-knowledge and the low-knowledge bets are based on vague probabilities whereas the chance lotteries have clear probabilities.

A noteworthy feature of Figure 6, which distinguishes it from the previous graphs, is that -- within each of the categories -- the preferences are essentially independent of P. Evidently, the competence effect is fully captured in this case by the contrast between the categories, hence the added knowledge implied by the judged probability has little or no effect on the choice among the bets.

Insert Figure 7 about here

Figure 7 presents the average calibration curves for Experiment 4, separately for the high- and low-knowledge categories. These graphs show that the judgments were generally overconfident: subjects' confidence exceeded their hit rate. Furthermore, the overconfidence was more pronounced in the high-knowledge category than in the low-knowledge category. As a consequence, the ordering of bets did not mirror judgmental accuracy. Summing across all triples, betting on the chance lottery would win 69% of the time, betting on the novice category would win 64% of the time and betting on the expert category would win only 60% of the time. By betting on the expert category, therefore, the subjects are losing, in effect, 15% of their

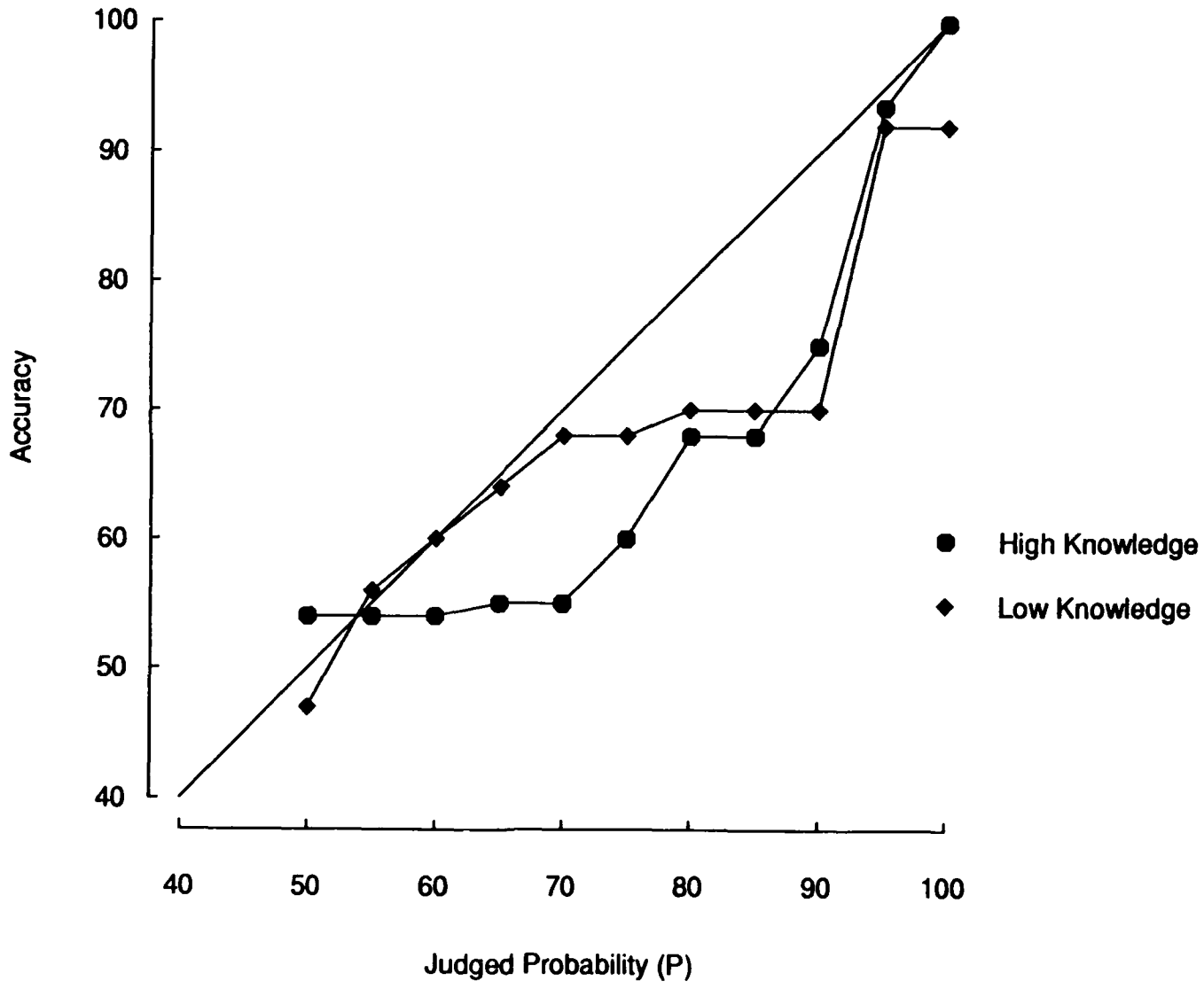


Figure 7. Calibration curves for high- and low-knowledge categories. (Experiment 4)

expected earning.

Experiment 5: Complementary Bets

The previous experiments demonstrated the presence of systematic discrepancy between judgments of probability and choice between bets, implied by the competence hypothesis. In the present experiment, we test the competence hypothesis in a different design, based on pricing data, which does not involve probability judgment and provides an estimate of the premium that subjects are paying, in effect, in order to bet on a high-knowledge item.

Sixty-eight Stanford students were instructed to state their cash equivalent (i.e., a reservation price) for each of twelve bets. They were told that one pair of bets would be chosen and a few students, selected at random, would play the bet for which they stated the higher cash equivalent. (For a discussion of this payoff scheme, see Tversky, Slovic & Kahneman, 1990.) All bets in this experiment offered a prize of $15 if a given proposition is true, and nothing otherwise. For example, half the subjects were asked to price the bet that pays $15 if the air distance between New York and San Francisco is more than 2500 miles, and nothing otherwise. The other half of the subjects were asked to price the complementary bet that pays $15 if the air distance between New York and San Francisco is less than 2500 miles, and nothing otherwise.

To investigate uncertainty preferences, we constructed matched pairs of high-knowledge and low-knowledge propositions. For example, we assumed that the subjects know more about the air distance between New York and San Francisco than about the air distance between Beijing and Bangkok. We also assumed that our respondents know more about the percentage of

undergraduate students who receive on-campus housing at Stanford than at the University of Nevada, Las Vegas. As before, we refer to the matched propositions as high-knowledge and low-knowledge items, respectively. Note that by varying the stated value of the uncertain quantity (e.g., air distance, percentage of students) one could vary the subjects' confidence in the validity of the proposition in question, independent of his or her general knowledge about the subject matter. Twelve pairs of matched problems were constructed, and each subject evaluated one of the four bets defined by each pair.

Summing across all twelve pairs of complementary propositions, subjects were willing to pay on average \$7.12 for the high-knowledge bets and only \$5.96 for the low-knowledge bets ($p < .01$). Thus, people were paying, in effect, a competence premium of nearly 20% in order to bet on the more familiar proposition. Furthermore, the average price for the (complementary) high-knowledge bets was greater than that for the low-knowledge bets in 11 out of 12 problems. Recall that, with additive probabilities and a linear utility function, the average cash equivalent should be \$7.50 for all the pairs. To control for the possibility that the subjective probabilities in the high-knowledge bets were more extreme than in the low-knowledge bets, we have restricted the analysis to bets in which the difference between the mean prices for the complementary propositions was less than \$3. This restriction eliminated four pairs of problems but it did not reduce the discrepancy: the average price for the high-knowledge bet was \$7.17 whereas the average price for the low-knowledge bet was \$5.78.

To devise a purely ordinal test of source independence, let H and $\overline{H}$ denote two complementary high-knowledge propositions, and let L and $\overline{L}$ denote the matching pair of complementary low-knowledge propositions. Suppose the decision maker prefers betting, say \$15, on H

over L, and at the same time prefers betting on $\overline{H}$ over $\overline{L}$. Such a pattern is inconsistent with expected utility theory because it violates the additivity of subjective probability. If, on the other hand, high-knowledge bets are preferred to low-knowledge bets, such a pattern is likely to arise, especially when H and L are roughly equiprobable; if H is much more probable than L, people will obviously bet on H rather than on L but in this case they will also bet on $\overline{L}$ rather than on $\overline{H}$. Because in the present experiment, the four propositions (H, $\overline{H}$, L, $\overline{L}$) were evaluated by four different groups of subjects, we employ a between-subject test of additivity. Let $M(H_i)$ be the median price for the high-knowledge proposition H_i, etc. The observed distribution of prices is said to violate additivity (in an ordinal sense) with respect to proposition i whenever

$$M(H_i) > M(L_i) \text{ and } M(\overline{H}_i) \geq M(L_i).$$

Five of the twelve pairs of matched problems used in the experiment exhibited this pattern, which is inconsistent with additive probability. For example, the median price for betting on the proposition that more than 85% of undergraduates at Stanford receive on-campus housing was \$7.50 and the median cash equivalent for betting on the complementary proposition was \$10. In contrast, the median cash equivalent for betting on the proposition that more than 70% of undergraduates at UNLV receive on-campus housing was \$3 and the median value for the complementary bet was \$7. The majority of respondents, therefore, were willing to pay more to bet on either side of a high-knowledge item than on either side of a low-knowledge item. For comparison, the cash equivalent for a coin flip to win \$15 was \$7, which falls between the average price of the high-knowledge propositions and the average price of the low-knowledge propositions. In accord with our previous findings regarding high-knowledge items, people value either

side of the uncertain proposition more than an even-chance bet, although the former is ambiguous and the latter is not.

The observation that the violations of additivity in five out of twelve problems favored the high-knowledge bet, and that none of the violations favored the low-knowledge bet indicates that the data cannot be attributed to random error. By chance, the two types of violations should be equally frequent. These findings confirm the competence hypothesis in a test that does not rely on judgments of probability or on the comparison of a judgment bet to a matched lottery. Hence, the present results cannot be attributed to peculiarities of the judgment process or to a regression bias in the matching of high- and low-knowledge items.

Discussion

The experiments reported in this paper establish a consistent and pervasive discrepancy between judgments of probability and choice between bets. The evidence shows that -- holding judged probability constant -- people prefer to bet in contexts in which they regard themselves knowledgeable or competent. Experiment 1 demonstrated that the preference for the judgment bet over the chance lottery increases with confidence, contrary to the hypothesis that people avoid ambiguity when the chances of success are moderate or high and seek ambiguity when the chances are low. Experiment 2 replicates this finding for future real-world events, and demonstrates a knowledge effect, independent of judged probability; Experiment 3 extends this finding to small probabilities. In Experiment 4, we sort subjects into their strong and weak areas and show that people like betting on their strong category, and dislike betting on their weak category; the chance bet is intermediate between the two. This finding cannot be explained by

ambiguity or by second-order probability because chance is unambiguous whereas judgmental probability is vague. Finally, Experiment 5 demonstrates uncertainty preferences in a pricing task that does not rely on probability matchings, and shows that people are paying, in effect, a 20% premium for betting on more familiar propositions.

The competence hypothesis can be used to explain other instances of uncertainty preferences reported in the literature, notably the preference for clear over vague probabilities in a chance setup, the preference to bet on the future over the past (Snyder & Rothbart, 1971; Brun & Teigen, 1989), the preference for skill over chance (Cohen & Hansel, 1959; Howell, 1971), and the enhancement of ambiguity aversion in the presence of knowledgeable others (Curley, Yates & Abrams, 1986). Our major empirical finding that, in their area of competence, people prefer to bet on their (vague) beliefs over a matched chance event demonstrates that the effect of knowledge or competence far outweight the contribution of ambiguity. An analysis of uncertainty preferences that is based on ambiguity and neglects the effect of general knowledge or competence, therefore, cannot provide an adequate account of the relation between belief and preference.

In Experiments 1-4 we used probability judgments to establish belief, and choice data to establish preference, and we have interpreted the choice-judgment discrepancy as a preferential phenomenon. This interpretation can be challenged on the ground that choice, not judgment, should be the proper basis for the assessment of subjective probability. According to this interpretation, the choice-judgment discrepancy is attributable to a judgmental bias, namely an underestimation of the probabilities of all high-knowledge items (whether they are high as in Experiment 2 or low as in Experiment 3), and an overestimation of the probabilities of all low-

knowledge items. This hypothesis, however, is not supported by the available evidence, because it implies less overconfidence for high-knowledge than for low-knowledge items, contrary to the fact (see Figure 7). Furthermore, this hypothesis cannot explain the results of Experiment 5, which demonstrates preferences for betting on high-knowledge items in a pricing task that does not involve probability judgment. Finally, judgments of probability cannot be dismissed as inconsequential because in the presence of a scoring rule, such as the one used in Experiment 4, these judgments represent another form of betting.

The distinction between preference and belief lies at the heart of decision theory. The standard conception of the theory of utility and subjective probability assume (i) that people's beliefs are consistent with an additive probability measure, (ii) that choice obeys the expectation principle, hence it gives rise to a (subjective) probability measure, and (iii) the subjective probability derived from choice is consistent with people's beliefs. It is important to realize that (i) and (ii) are logically independent. A person's beliefs may be described by the ordering of events according to their perceived likelihood, and this ordering may or may not be compatible with an additive probability measure. At the same time, the preferences of that person may or may not satisfy the expectation principle. Allais' problem, for example, and the subsequent demonstrations of the non-linearity of preferences under risk violate (ii) but not (i). Indeed, many authors have introduced non-additive decision weights to accommodate the observed violations of expected utility theory. These decision weights, derived from one's preferences, need not agree with one's belief. A person may believe that the probability of drawing the ace of spades from a well-shuffled deck is 1/52, yet he or she may give this event a higher weight when pricing a bet that is contingent on that event. Similarly, Ellsberg's example does not establish that people

regard the clear event as more probable than the corresponding vague event although this possibility cannot be ruled out; it only shows that people prefer to bet on the clear event. Unfortunately, the term *subjective probability* is used in the literature to describe the decision weights derived from preference as well as a measure of belief inferred from a subjective ordering of events. Under expected utility theory, the two measures coincide, hence the term subjective probability. As we go beyond this theory, however, it is essential to distinguish between decision weights and probability judgments.

The distinction between preference and belief is particularly important for the interpretation of ambiguity effects. There is evidence that, when the probability of winning is small or when the probability of losing is high, people seem to prefer ambiguity to clarity (Gärdenfors & Sahlin, 1982; Einhorn & Hogarth, 1985; Hogarth & Kunreuther, 1989). These observations suggest that people overweight vague events with low probabilities and underweight vague events with high probabilities, as implied by the regression hypothesis discussed earlier. To interpret these data, however, it is essential to determine whether they reflect uncertainty preferences or variations of belief. Einhorn and Hogarth (1985), for example, gave subjects numerical probability estimates and manipulated ambiguity by varying the stated reliability of the estimate. But in order to interpret people's willingness to bet on these events, we must establish that they are perceived as equiprobable. After all, information about the reliability of an estimate could change the estimate itself, not only the expected error. To investigate this question, we replicated the manipulation of ambiguity used by Einhorn and Hogarth (1985); one group of subjects (N=52) received the following information.

> Imagine that you head a department in a large insurance company. The owner of a small business comes to you seeking insurance against $1000 loss which could result from claims concerning a defective product. Based on the understanding of the manufacturing process, the reliabilities of the machines used, and the evidence contained in the business records, independent observers have stated that the probability of a defective product is .01, and that you could feel confident about the estimate.

A second group of subjects (N=50) received the same information, except that the last phrase "you could feel confident about the estimate" was replaced by "you could experience considerable uncertainty about the estimate," as in the original study. All subjects were now asked "given this information, what is your best guess about the probability of experiencing a loss (Check one)."

Above .01 _____

Below .01 _____

Exactly .01 _____

Insert Table 4 about here

The two groups were also asked to evaluate a second case in which the stated probability of a loss was .90. Table 4 presents the percentage of subjects in the two groups who chose each of the three responses for the stated values of .01 and .90. The distribution of responses under the high-reliability condition were roughly symmetric around the stated value. In contrast, the distribution of responses for the low-reliability condition is markedly asymmetric. The interaction between reliability (high-low) and direction (above-below) is highly significant ($p < .01$).

Table 4. Subjective assessments of stated probabilities of .01 and .90 under high- and low-reliability conditions. The entries are the percentage of subjects who chose each of the three responses.

Stated Value	Response	High Reliability	Low Reliability
.01	Above .01	46	80
	Exactly .01	15	6
	Below .01	39	14
.90	Above .90	42	26
	Exactly .90	23	12
	Below .90	35	62

Telling subjects that they "could experience considerable uncertainty about the estimate" produces a regressive shift: the majority of subjects expect the probability of loss to be above .01 in the first problem and below .90 in the second. As a consequence, the bet on the reliable estimate should be more attractive than the bet on the unreliable estimate when the probability of loss is low (e.g., .01), and the opposite should hold when the probability of loss is high (e.g., .90). This is exactly the pattern of preference observed by Einhorn and Hogarth (1985) and by Hogarth and Kunreuther (1989). This pattern, however, need not reflect either ambiguity seeking or ambiguity aversion; it may be simply due to the fact that people's best guess about an extreme estimate with low reliability is generally regressive. This belief is not necessarily unreasonable, and it can even be rationalized by a suitable prior distribution. The experimental results obtained by manipulating the reliability of a stated estimate, therefore, may reflect differences in belief rather than uncertainty preferences.

The problem of separating beliefs from preferences arises in other manipulations of ambiguity as well. Unlike Ellsberg's comparison of the 50/50 box with the unknown box in which symmetry essentially precludes a bias in one direction or another, the manipulation of ambiguity in asymmetric problems could confound belief and preference, as demonstrated in an unpublished study by Parayre and Kahneman conducted in 1985.

Insert Table 5 about here

Table 5. (Data from Parayre and Kahneman). Percentage of subjects who selected the sharp event and the vague event in judgment of likelihood and in direct choice. The sum of the two values in each condition is less than 100%; the remaining responses expressed equivalence. Significant differences at the .05 level are denoted by an * near the larger value. In the likelihood rating task, the low values were .05 and [0,.1].

Probability (Win/lose)		Judgment	Choice	
			Win $100	Lose $100
Low	.075	26	12	66*
	[0,.15]	55*	74*	12
Medium	.5	37	60*	60*
	[0,1]	25	26	21
High	.9	55*	50	22
	[.8,1]	22	34	47*

The investigators compared a sharp event, defined by the proportion of red balls in the box, with a vague event defined by the range of balls of the designated color. For example, a vague event was generated by informing the subject that the percentage of red balls could be anywhere between 80% and 100%, compared with 90% for the sharp event. Table 5 presents both choice and judgment data for three probability levels: low, medium and high. The choice data show that subjects preferred the ambiguous box when the probability of winning was low and when the probability of losing was high, as observed by other investigators. The novel feature of the Parayre and Kahneman experiment lies in the use of probability judgment. Using a perceptually-based (non-numerical) rating scale, these investigators showed that the judged probabilities were regressive with respect to the stated values. That is, the vague low-probability event was judged as more probable than the clear event, and the vague high-probability event was judged as less probable than the sharp event. There was no significant difference in the likelihood rating of the medium probability. These results, like the previous finding, demonstrate that the preference for the ambiguous event, observed at the low end for positive bets and at the high end for negative bets, reflect variations in the perception of probability rather than willingness or unwillingness to bet on uncertain events.

The picture that emerges from these results is much more complicated than the standard theory. According to the classical analysis, people have beliefs that are consistent with an additive probability measure, and they invariably select the prospect with the highest expected utility, computed relative to their subjective probability. It is evident by now that this conception is descriptively inadequate. Allais was the first to show that people do not maximize expected utility even in a chance context, and Ellsberg has shown that risk (in the sense of known probabili-

ties) and uncertainty are not treated alike, although his result can be interpreted either as a judgment effect that entails non-additive probabilities, or as a preference to bet on chance over ignorance. In accord with the latter interpretation, the present results demonstrate that people like to bet on their judgment when they consider themselves knowledgeable or competent, and that they avoid betting on their judgment when they feel uninformed or ignorant. This result challenges not only expected utility theory; it also challenges the very idea of using preferences to infer beliefs. For if people's willingness to act depends not only on the degree of uncertainty (and the precision with which it is measured) but also on one's general knowledge of the domain and his or her sense of competence concerning a particular proposition, it is exceedingly difficult, if not impossible to derive underlying beliefs from observed preferences.

References

Barlow, R. E., Bartholomew, Brimmer, & Brink (1972). *Statistical inference under order restrictions: The theory and application of isotonic regression.* New York: J. Wiley.

Brun, W., & Teigen, K. H. (in press). Prediction and postdiction preferences in guessing. *Journal of Behavioral Decision Making.*

Budescu, D. V., Weinberg, S., & Wallsten, T. S. (1988). Decisions based on numerically and verbally expressed uncertainties. *Journal of Experimental Psychology: Human Perception and Performance, 14* (2), 281-294.

Cohen, J., & Hansel, C. E. M. (1959). Preferences for different combinations of chance and skill in gambling. *Nature, 183,* 841-843.

Curley, S. P., & Yates, J. F. (1989). An empirical evaluation of descriptive models of ambiguity reactions in choice situations. *Journal of Mathematical Psychology, 33,* 397-427.

Curley, S. P., Yates, J. F., & Abrams, R. A. (1986). Psychological sources of ambiguity avoidance. *Organizational Behavior and Human Decision Processes, 38,* 230-256.

Einhorn, H. J., & Hogarth, R. M. (1985). Ambiguity and uncertainty in probabilistic inference. *Psychological Review, 93,* 433-461.

Ellsberg, D. (1961). Risk, ambiguity, and the Savage axioms. *Quarterly Journal of Economics, 75,* 643-669.

Ellsberg, D. (1963). Risk, ambiguity, and the Savage axioms: Reply. *Quarterly Journal of Economics, 77,* 336,342.

Fellner, W. (1961). Distortion of subjective probabilities as a reaction to uncertainty. *Quarterly Journal of Economics, 75,* 670-689.

Frisch, D., & Baron, J. (1988). Ambiguity and rationality. *Journal of Behavioral Decision Making, 1,* 149-157.

Gärdenfors, P., & Sahlin, N-E. (1982). Unreliable probabilities, risk taking, and decision making. *Synthese, 53* (3), 361-386.

Hogarth, R. M., & Kunreuther, H. (1988). *Pricing insurance and warranties: Ambiguity and correlated risks*. Unpublished manuscript, University of Chicago and University of Pennsylvania.

Hogarth, R. M., & Kunreuther, H. (1989). Risk, ambiguity, and insurance. *Journal of Risk and Uncertainty, 2,* 5-35.

Howell, W. C. (1971). Uncertainty from internal and external sources: A clear case of overconfidence. *Journal of Experimental Psychology, 89* (2), 240-243.

Kahneman, D., and Tversky, A. (1984). Choices, values and frames. *American Psychologist, 39,* 341-350.

Lichtenstein, S., Fischhoff, B., & Phillips, L. D. (1982). Calibration of probabilities: The state of the art to 1980. In D. Kahneman, P. Slovic & A. Tversky, A. (Eds.), *Judgment under uncertainty: Heuristics and biases*. New York: Cambridge University Press.

March, J. G., & Shapira, Z. (1987). Managerial perspectives on risk and risk taking. *Management Science, 33* (11), 1404-1418.

Raiffa, H. (1961). Risk, ambiguity, and the Savage axioms: Comment. *Quarterly Journal of Economics, 75,* 690-694.

Ramsey, F. P. (1931). Truth and probability. In F. P. Ramsey, *The foundations of mathematics and other logical essays*. NY: Harcourt, Brace and Co.

Rothbart, M., & Snyder, M. (1970). Confidence in the prediction and postdiction of an uncertain outcome. *Canadian Journal of Behavioral Science, 2* (1), 38-43.

Savage, L. J. (1954). *The foundations of statistics*. NY: Wiley.

Tversky, A., & Kahneman, D. (1989). *Loss Aversion in risky and riskless choice*. Unpublished manuscript, Stanford University and University of California at Berkeley.

Figure Captions

Figure 1. Five contrasting predictions of the results of an uncertainty preference experiment.

Figure 2. Percentage of choices (C) that favor a judgment bet over a matched lottery as a function of judged probability (P) in Experiment 1.

Figure 3. Calibration curve for Experiment 1.

Figure 4. Percentage of choices (C) that favor a judgment bet over a matched lottery as a function c judged probability (P), for high- and low- knowledge items in the football prediction task (Experiment 2).

Figure 5. Percentage of choices (C) that favor a judgment bet over matched lottery as a function of judged probability (P), for high- and low- knowledge items in Experiment 2 (Election data).

Figure 6. Ranking data for high knowledge, low knowledge and chance bets as a function of P in Experiment 4.

Figure 7. Calibration curves for high- and low-knowledge categories in Experiment 4.

To appear in R. M. Hogarth (Ed.), Insights in decision making: Theory and applications. (A tribute to the late Hillel J. Einhorn.)

Compatibility Effects in Judgment and Choice

Paul Slovic, Decision Research and The University of Oregon

Dale Griffin, University of Waterloo

Amos Tversky, Stanford University

Support for this research was provided by Grant 89-0064 of the Air Force Office of Scientific Research to Stanford University, and by NSF Grant No. SES 8712145 to Decision Research. This chapter has benefited from the comments of Robyn Dawes, Gregory Fischer, Robin Hogarth, Eric Johnson, and Daniel Kahneman.

Abstract

We investigate the hypothesis that the weight of a stimulus attribute is enhanced by its compatibility with the response mode. The first section demonstrates compatibility effects in predictions of market value (Study 1) and course grades (Study 2). In each case, the weight of a stimulus attribute is greater when it matches the response scale than when it does not. The second section applies the compatibility principle to the study of choice, and investigates the hypothesis that preference reversals are caused by the fact that payoffs are weighted more heavily in pricing than in choice, as implied by compatibility. This account is supported in experiments on risky choice (Studies 3 and 5), and on time preferences (Study 4). Theoretical and practical implications of the compatibility hypothesis are discussed in the last section.

One of the main ideas that has emerged from behavioral decision research in the last two decades is a constructive conception of judgment and choice. According to this view, preferences and beliefs are actually constructed--not merely revealed--in the elicitation process. This conception is entailed by findings that normatively equivalent methods of elicitation often give rise to systematically different responses (see, e.g., Slovic, Fischhoff & Lichtenstein, 1982; Tversky, Sattath, & Slovic, 1988). To account for these data within a constructive framework, we seek explanatory principles that relate the characteristics of the task to the attributes of the objects under study. One such notion is the compatibility hypothesis, which states that the weight of a stimulus attribute is enhanced by its compatibility with the response.

The rationale for this hypothesis is twofold. First, non-compatibility between the input and the output requires additional mental operations, which often increase effort and error and may reduce impact. Second, a response mode may prime or focus attention on the compatible features of the stimulus. Common features, for example, are weighted more heavily in judgments of similarity than in judgments of dissimilarity, whereas distinctive features are weighted more heavily in judgments of dissimilarity (Tversky, 1977). Consequently, entities with many common features and many distinctive features (e.g., East Germany and West Germany) are judged as both more similar to each other and as more different from each other than entities with relatively fewer common features and fewer distinctive features (e.g., Sri Lanka and Nepal).

The significance of the compatibility between input and output has long been recognized by students of human performance. Engineering psychologists have discovered that responses to visual displays of information, such as an instrument panel, will be faster and more accurate if

the response structure is compatible with the arrangement of the stimuli (Fitts & Seeger, 1953; Wickens, 1984). For example, the response to a pair of lights will be faster and more accurate if the left light is assigned to the left key and the right light to the right key. Similarly, a square array of four burners on a stove is easier to control with a matching square array of knobs than with a linear array. The concept of compatibility has been extended beyond spatial organization. The reaction time to a stimulus light is faster with a pointing response than with a vocal response, but the vocal response is faster than pointing if the stimulus is presented in an auditory mode (Brainard, Irby, Fitts, and Alluisi 1962).

The present chapter investigates the role of compatibility in judgment and choice. As in the study of perceptual-motor performance, we do not have an independent procedure for assessing the compatibility between stimulus elements and response modes. This hinders the development of a general theory, but it does not render the concept meaningless or circular, provided compatibility can be experimentally manipulated. For example, it seems reasonable to assume that a turn signal in which a left movement indicates a left turn and a right movement indicates a right turn is more compatible than the opposite design. By comparing people's performance with the two turn signals, it is possible to test whether the more compatible design yields better performance. Similarly, it seems reasonable to assume that the monetary payoffs of a bet are more compatible with pricing than with choice, because both the payoffs and the prices are expressed in dollars. By comparing choice and pricing, therefore, we can test the hypothesis that the payoffs of a bet loom larger in pricing than in choice.

The research described in this chapter employs the notion of compatibility as a guiding principle that is translated into specific experimental hypotheses. In the first section we demonstrate compatibility effects in studies of prediction. The next section applies the compatibility hypothesis to the analysis of preference reversals in both risky and riskless choice. Theoretical and practical implications of the findings are addressed in the final section.

Prediction

Study 1: Prediction of Market Value

In all the following studies, subjects were either undergraduate students at Stanford University participating for course credit, or students at the University of Oregon who responded to an ad in the student newspaper and were paid for their participation. In our first study, seventy-seven Stanford students were presented with a list of twelve well-known U.S. companies taken from the 1987 ***Business Week*** Top 100. For each company, students were given two items of information: i) 1986 ***market value*** (i.e., the total value of the outstanding shares in billions of dollars), and ii) 1987 ***profit standing*** (i.e., the rank of the company in terms of its 1987 earnings among the Top 100); see Table 1. Half of the subjects were asked to predict 1987 market value (in billions of dollars). They were informed that the highest market value in 1987 was 68.2 billion dollars and the lowest (among the Top 100) was 5.1 billion dollars, so their predictions should fall within that range. The remaining subjects were asked to predict each company's rank (from 1 to 100) in market value for 1987. Thus, both groups of subjects received identical information and predicted the same criterion, using a different response scale. Although the two response scales differ in units (dollar versus rank) and direction (low rank means high

market value), the two dependent variables should yield the same ordering of the twelve companies. To encourage careful consideration, a $75 prize was offered for the person whose predictions most nearly matched the actual values. The mean predicted values for each group are presented in Table 1 along with the actual values.

Insert Table 1 about here

The compatibility hypothesis states that a predictor will be weighted more heavily when it matches the response scale than when it does not. That is, 1986 market value in dollars should be weighted more heavily by the subjects who predict in dollars than by those who predict in rank. By the same token, 1987 profit rank should be weighted more heavily by the subjects who predict in rank than by those who predict in dollars. To investigate this hypothesis, we (i) correlated the criteria with the predictors, (ii) estimated the relative weights of the two predictors, and (iii) devised a statistical test based on reversals of order.

The product-moment correlations of d with D and R were .93 and .77, respectively, whereas the correlations of r with D and R were .74 and .94. Thus, the correlation between the matched variables was higher than that between the nonmatched variables. It is instructive to examine the compatibility effect in terms of the relative weights of the two predictors in a multiple regression equation. These values can be computed directly, or derived from the correlations between the predictors and the criterion together with the correlation between the predictors. (To make the regression weights positive, the ranking order was reversed). The multiple regres-

Table 1

Financial information for the twelve companies used to test the compatibility hypothesis with the respective mean predictions (actual outcome values in parenthesis).

		Predictors		*Criteria*	
		D	R	d	r
#	Company	1986 Market Value in billions	1987 Profit Rank (1 to 100)	1987 Market Value in billions	1987 Market Rank (1 to 100)
1	Chevron Corp.	$18.0	26	$21.3 (16.2)	30 (15)
2	H. J. Heinz	$6.2	75	$7.3 (5.6)	70 (84)
3	Coca-Cola	$18.1	31	$21.6 (14.8)	31 (17)
4	Westinghouse	$9.3	36	$12.9 (7.4)	44 (51)
5	Dow Chemical	$15.5	16	$20.5 (16.9)	26 (13)
6	Xerox	$7.1	54	$9.5 (5.7)	53 (82)
7	Chrysler	$8.2	12	$15.5 (5.5)	32 (90)
8	Kraft	$8.4	74	$9.0 (7.3)	64 (53)
9	Hewlett-Packard	$14.7	39	$17.4 (15.5)	42 (16)
10	Procter & Gamble	$15.6	63	$16.3 (13.9)	47 (25)
11	Kodak	$16.9	20	$20.9 (13.7)	27 (26)
12	Johnson & Johnson	$15.5	35	$18.2 (14.7)	36 (18)

sions for both dollars and ranks fit the average data very well with multiple correlations of .99. Let d_i and r_i denote the mean observed predictions of 1987 dollar value and rank, respectively, for a company whose 1986 dollar value is D_i and whose 1987 profit rank is R_i. The multiple regression equations, then, take the form

$$d_i = \alpha_d D_i + \beta_d R_i$$

$$r_i = \alpha_r D_i + \beta_r R_i,$$

when the independent variables are expressed in standardized units. Thus, α_d and α_r are the regression weights for the 1986 market value (D_i) estimated respectively from the predicted dollars and ranks. Similarly, β_d and β_r are the corresponding weights for the second predictor, 1987 profit rank. The relative weights for the first predictor in each of the two response modes are

$$A_d = \alpha_d/(\alpha_d+\beta_d)$$

and

$$A_r = \alpha_r/(\alpha_r+\beta_r).$$

These values measure the relative contribution of D_i in the prediction of dollars and rank, respectively. If the weighting of the dimensions is independent of the response scale, A_d and A_r are expected to be equal, except for minor perturbations due to a nonlinear relation between d and r. As we shall argue next, the compatibility hypothesis implies $A_d > A_r$. Note that A_d is the relative weight of the 1986 market value in dollars, estimated from the prediction of dollars, whereas A_r is the relative weight of the same variable estimated from the prediction of rank. The first index reflects the impact of D_i in a compatible condition (i.e., when the predictions are made in dollars), while A_r reflects the impact of D_i in the less compatible condition (i.e., when the predictions are made in ranks). If the compatibility between the predictor and the criterion enhances

the weight of that variable then A_d should exceed A_r.

The values estimated from the regression equations were A_d=.64 and A_r=.32, in accord with the compatibility hypothesis. Thus, D_i was weighted more than R_i in the prediction of dollars, whereas R_i was weighted more than D_i in the prediction of rank. Moreover, each predictor was weighted about twice as much in the compatible condition than in the non-compatible condition. When interpreting the relative weights, here and in later studies, we should keep in mind (i) that they are based on aggregate data, (ii) that the predictors (D and R) are correlated, and (iii) that the relation between the two criteria (d and r) should be monotone but not necessarily linear. Although these factors do not account for the discrepancy between A_d and A_r, it is desirable to obtain a purely ordinal test of the compatibility hypothesis within the data of each subject that is not open to these objections. The following analysis of order reversals provides a basis for such a test.

The change in the relative weights induced by the response mode could produce reversals in the order of the predictions. In the present study, there were 21 pairs of companies (i,j) in which $D_i > D_j$ and $R_j > R_i$. If D is weighted more heavily than R in the subject's prediction of dollars, and R is weighted more heavily than D in the subject's prediction of rank, we would expect $d_i > d_j$ and $r_j > r_i$. The data confirmed this hypothesis. Subjects who predicted dollars favored the company with the higher D 72% of the time, whereas subjects who predicted rank favored the company with the higher D only 39% of the time. (Ties were excluded from this analysis.) This difference is highly significant ($p < .001$). Note that the subjects did not directly compare the companies; the ordering was inferred from their predictions.

Insert Figure 1 about here

Figure 1 provides a graphical summary of the stimuli and the data. Each of the twelve companies is represented as a point in the D × R plane. Each regression equation defines a set of parallel equal-value lines. The points on any given line are the values of the two predictors that give rise to the same predicted value of the criterion. For the prediction of dollar, for instance, each equal-value line is the set of points for which $\alpha_d D_i + \beta_d R_i$ is a constant. The two prediction lines (d and r) are perpendicular to the equal-value lines for the two criteria. Hence, the predicted order of the companies is given by the order of their projections, denoted by notches. The slopes of the prediction lines are the weight ratios, α_d/β_d and α_r/β_r, of D to R, estimated from d and r, respectively. It is evident from the figure that, in accord with the compatibility hypothesis, the two criteria induced different orders of the twelve companies. For example, the predicted market value of Chevron (#1) is higher than that of Dow Chemical (#5), but the latter is assigned a higher rank than the former.

Study 2: Prediction of academic performance

Our second test of the compatibility hypothesis involves the prediction of a student's grade in a course. Two hundred and fifty-eight subjects from the University of Oregon predicted the performance of 10 target students in a History course on the basis of the students' performance in two other courses: English Literature and Philosophy. For each of the 10 targets, the subjects were given a letter grade (from A+ to D) in one course, and a class rank (from 1 to 100)

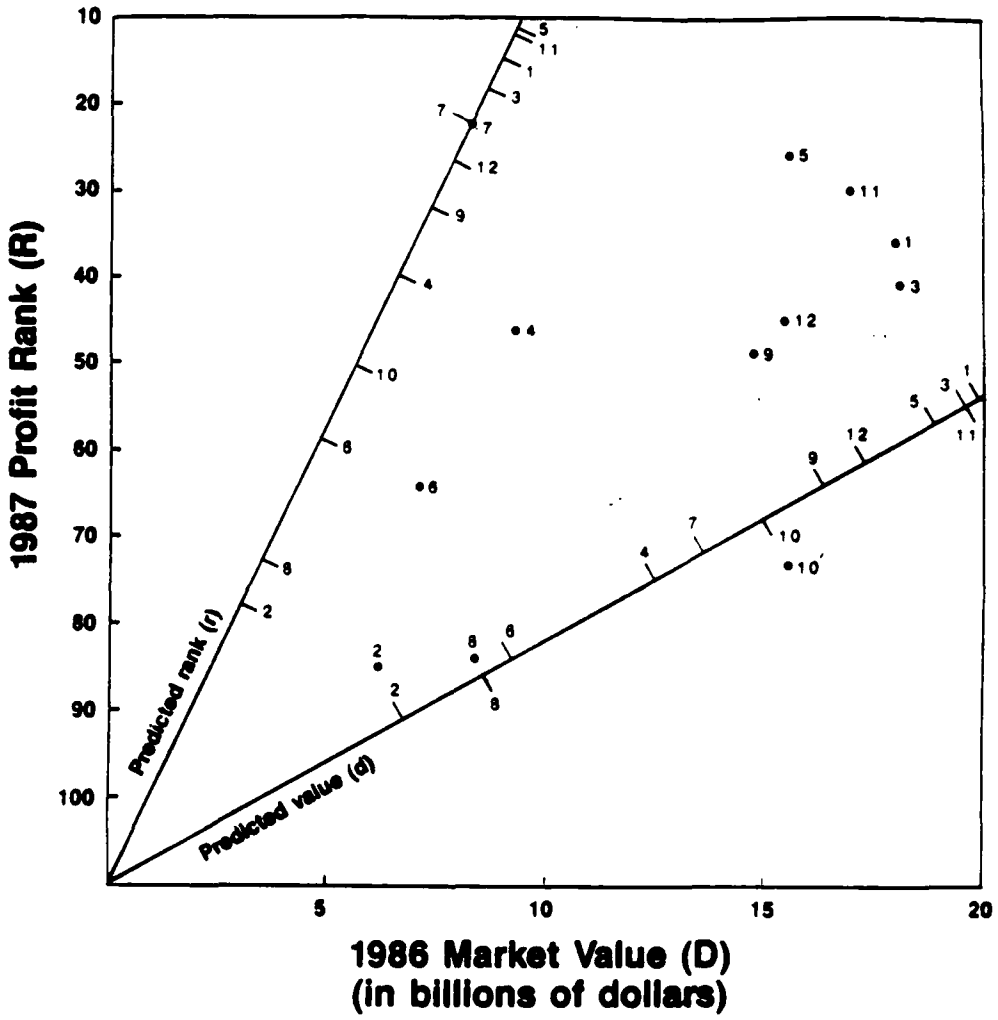


Figure 1. A graphical summary of Study 1. The dots represent the twelve companies. The slopes of the d and r lines correspond to the weight ratios, α_d/β_d and α_r/β_r, of D to R, in the two prediction tasks.

in the other course. One-half of the subjects predicted the students' grade in History whereas the other half predicted the students' class rank in History. Each of the four combinations of performance measures (grade/rank) and courses (Literature/Philosophy) was presented to a different group of subjects. The description of the 10 hypothetical students is presented in Table 2 along with the mean predictions of grade and rank, rounded to the nearest unit.

Insert Table 2 about here

The compatibility hypothesis implies that a given predictor (e.g., grade in Philosophy) will be given more weight when the criterion is expressed on the same scale (e.g., grade in History) than when it is expressed on a different scale (e.g., rank in History). The relative weight of grades to ranks, then, will be higher in the group that predicts grades than in the group that predicts ranks.

As in the previous study, we first correlated the criteria with the predictors. The (zero-order) correlations of g with G and R were .83 and .82, respectively, whereas the correlations of r with G and R were .70 and .91, in accord with the compatibility hypothesis. We next regressed the mean predictions of grades and ranks (displayed in Table 2) onto the two predictors. The letter grades were coded D=1, C-=2 ... A+=10. (To make the the regression weights positive, the ranking order was reversed). The multiple regressions for both grades and ranks fit the average data very well with multiple correlations of .99. Let g_i and r_i denote the mean observed predictions of grade and rank, respectively, for a student with a grade G_i in one course and a rank R_i in

Table 2

Academic performance of the ten hypothetical students used to test the compatibility hypothesis with the respective mean predictions.

	Predictors		*Criteria*	
	G	R	g	r
Student	Grade in Class 1 (A+ to D)	Rank in Class 2 (1 to 100)	Predicted Grade	Predicted Rank
1	B+	66th	C+	48
2	D	93rd	D	87
3	A	45th	B	33
4	C+	34th	B-	40
5	A+	6th	A	11
6	C-	54th	C	54
7	B	59th	B-	49
8	A-	72nd	B-	48
9	C	28th	B-	35
10	B-	41st	B-	38

the other course. There was no significant interaction between the scale (rank/grade) and the course (Literature/Philosophy), therefore, the data for the two courses were pooled. The multiple regression equations, then, take the form

$$g_i = \alpha_g G_i + \beta_g R_i$$

$$r_i = \alpha_r G_i + \beta_r R_i,$$

when the independent variables are expressed in standardized units. Thus, α_g and α_r are the regression weights for the grades (G_i) estimated respectively from the predicted grades and ranks. Similarly, β_g and β_r are the corresponding weights for the second predictor, class rank. The relative weights for the first predictor in each of the two response modes are

$$A_g = \alpha_g/(\alpha_g+\beta_g)$$

and

$$A_r = \alpha_r/(\alpha_r+\beta_r).$$

These values measure the relative contribution of G_i in the prediction of grade and rank, respectively. Because the grades and ranks are monotonically related, A_g and A_r should be approximately equal if the weighting of the dimensions is independent of the response scale. However, if the match between the predictor and the criterion enhances the weight of the more compatible predictor, then A_g should exceed A_r.

The values estimated from the regression equations were $A_g = .51$ and $A_r = .40$, in accord with the compatibility hypothesis. Thus, grade in Philosophy was weighted more heavily in the prediction of grade in History than in the prediction of rank in History. Similarly, rank in Philosophy was weighted more heavily in the prediction of rank in History than in the prediction of grade in History.

To obtain an ordinal test of the compatibility hypothesis within the data of each subject, we analyzed the reversals of order induced by the change in weights. There were 21 pairs of students (i,j) in which $G_i > G_j$ and $R_j > R_i$. If G is weighted more heavily than R in the prediction of grades, and R is weighted more heavily than G in the prediction of rank, we would expect $g_i > g_j$ and $r_j > r_i$. Indeed, subjects who predicted grades favored the student with the higher G 58% of the time, whereas subjects who predicted rank favored the student with the higher G only 42% of the time. (Ties were excluded from this analysis.) This difference is statistically significant ($p < .001$). Recall that subjects did not compare students directly; the ordering was inferred from their predictions. Figure 2 provides a graphical representation of these data.

Insert Figure 2 about here

The compatibility effects observed in the previous two studies may be mediated by a process of anchoring and adjustment. Subjects may use the score on the compatible variable (the attribute which matches the criterion) as an anchor, and then adjust this number upward or downward according to the value of the non-compatible variable. Because adjustments of an anchor are generally insufficient (Slovic & Lichtenstein, 1971; Tversky & Kahneman, 1974) the compatible attribute would be overweighted. An anchoring and adjustment process, therefore, provides a natural mechanism for generating compatibility effects. To test whether compatibility effects occur in the absence of anchoring, we replaced the prediction task described above with a choice task in which the subject is no longer required to make a numerical prediction that

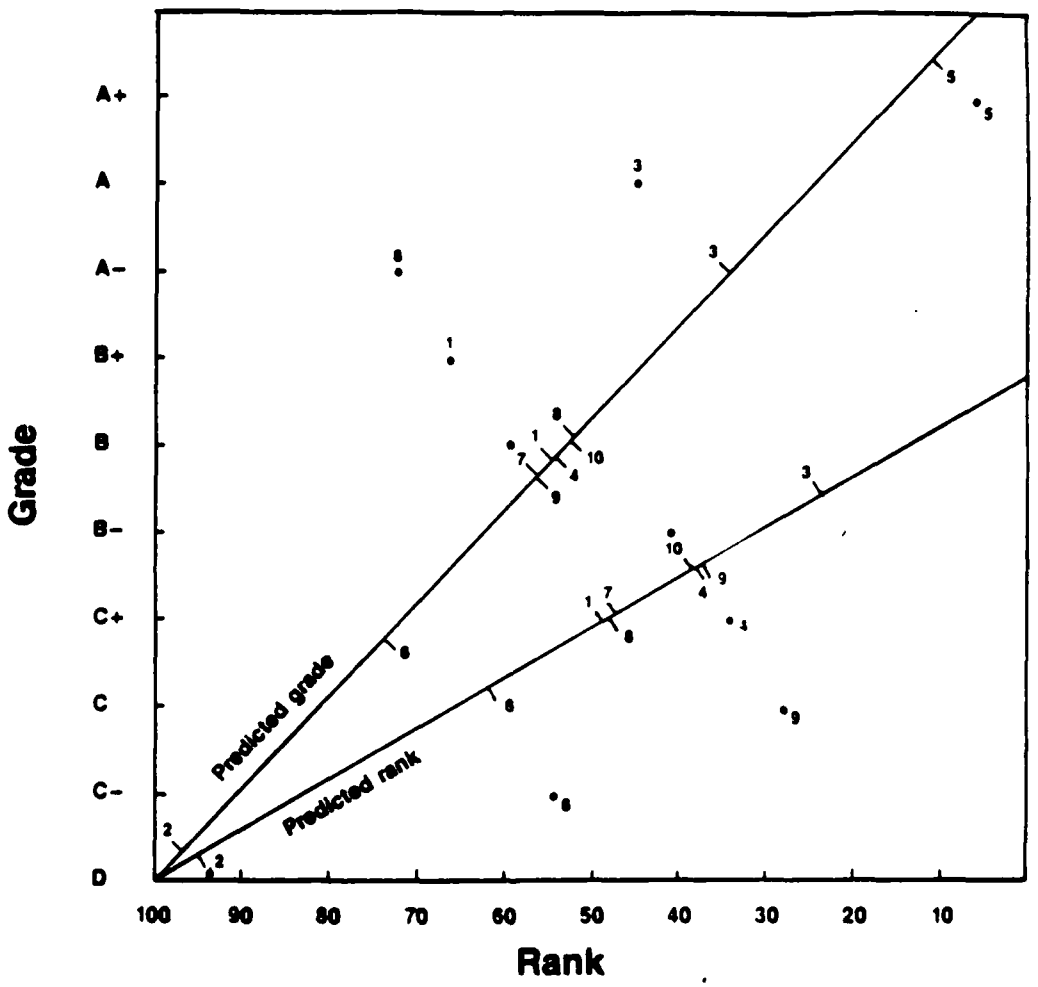


Figure 2. A graphical summary of Study 2. The dots represent the ten students. The slopes of the grade and rank lines correspond to the weight ratios, α_g/β_g and α_r/β_r, of grades to ranks in the two prediction tasks.

would invoke anchoring. The following study, then, investigates the compatibility hypothesis in a context in which anchoring and adjustment are unlikely to play a significant role.

Seventy-eight Stanford undergraduates were presented with 20 pairs of students taken from the list given in Table 2. In each pair, one student had a higher grade while the other had a higher rank. Half of the subjects were asked to predict, for each pair, which student would achieve a higher grade in History whereas the remaining subjects were asked to predict, for each pair, which student would achieve a higher rank in History. Because both groups were only asked to predict which of two students would do better in History, without making a numerical prediction, their tasks were virtually identical.

Nevertheless, the student with the higher grade was selected 56% of the time by the grade group and only 49% of the time by the rank group ($p < .05$), indicating that the compatibility effect is present even in a choice task that does not require a quantitative response and is, therefore, unlikely to involve an adjustment of a numerical anchor. The strategy of anchoring and adjustment, however, probably contributes to the compatibility effect observed in numerical predictions.

Preference

The previous section investigated compatibility effects in prediction and judgment. The present section is concerned with the role of compatibility in decision making in general, and preference reversals in particular. A reversal of preference is a pattern of choices in which normatively equivalent elicitation procedures give rise to inconsistent preferences. A well-known example of preference reversal was discovered by Lichtenstein and Slovic (1971; see also Slovic

& Lichtenstein, 1968). This phenomenon involves pairs of bets with comparable expected values: an H bet that offers a high probability of winning a relatively small amount of money (e.g., 35/36 chances to win $4) and an L bet that offers a low probability of winning a moderate amount of money (e.g., 11/36 chances to win $16). When offered a choice between such bets, most people choose the H bet over the L bet, but when asked to state the lowest selling price of each bet, the majority state a higher price for the L bet than for the H bet. In general, about half the subjects state prices that are inconsistent with their choices, thereby exhibiting a preference reversal, or PR for short. This pattern of preferences, which violates the standard theory of rational choice, has been observed in numerous experiments, including a study conducted on the floor of a Las Vegas casino (Lichtenstein & Slovic, 1973), and it persists even in the presence of monetary incentives designed to promote consistent responses (see, e.g., Grether & Plott, 1979; Slovic & Lichtenstein, 1983).

Let C_H and C_L denote, respectively, the cash equivalent (or minimum selling price) of the H bet and L bet, and let $>^*$ and $\sim$ denote strict preference and indifference, respectively. In this notation, PR is expressed as $H >^* L$, and $C_L > C_H$. Note that $>^*$ refers to preference between options, whereas $>$ refers to the ordering of cash amounts. (Naturally, $X > Y$ implies $X >^* Y$, that is, more money is preferred to less.)

It can be shown that PR violates either transitivity or procedure invariance, and possibly both (Tversky, Slovic & Kahneman, 1989). Procedure invariance states that choice and pricing yield the same ordering of options, that is, a bet B is preferred to a cash amount X if and only if the cash equivalent of B, C_B, exceeds X. In particular, $C_B = X$ whenever the decision maker is indifferent between playing the bet B and receiving the cash amount X. If procedure invariance

holds, PR reduces to an intransitivity of the form

$$C_H \approx H >^* L \approx C_L >^* C_H.$$

On the other hand, $>^*$ may be transitive, in which case PR violates procedure invariance. This violation can be produced by either

(i) overpricing of L (i.e., $C_L >^* L$), or

(ii) underpricing of H (i.e., $H >^* C_H$).

It follows from this analysis that PR may be caused either by the intransitivity of $>^*$ or by a failure of procedure invariance that gives rise to a choice-pricing discrepancy. To investigate these possibilities, Tversky et al. (1989) extended the traditional design by including, in addition to the bets H and L, a cash amount X that is compared to both. By focusing on all cases in which $C_L > X > C_H$, it is possible to diagnose all PR patterns according to whether they imply an intransitive choice, an overpricing of L, an underpricing of H, or both overpricing of L and underpricing of H. Tversky et al (1989) applied this analysis to an extensive study of preference reversals, using 18 triples (H, L, X) that covered a wide range of probabilities and payoffs. The diagnostic analysis of the observed response patterns showed that the most important determinant of PR was the overpricing of L. Intransitive choice and the underpricing of H played a relatively minor role, each accounting for less than 10% of the total number of reversals.

The compatibility hypothesis offers a simple explanation for the overpricing of L bets. Because the selling price of a bet is expressed in dollars, we expect that the payoffs, which are expressed in the same units, will be weighted more heavily in pricing than in choice. To test this

hypothesis, Tversky et al (1989) have employed a contingent weighting model in which the relative weight of an attribute varies with the method of elicitation. This analysis differs from the regression analysis discussed in the previous section in two important respects. First, the two attributes of a simple gamble, probability and payoff, combine multiplicatively rather than additively. Consequently, the multiple regression analysis was applied to the logarithms of the probabilities and the payoffs. Second, the analysis uses only the ordering of the bets by price and by choice. Specifically, assume that a bet $B = (P,X)$ is chosen over $B' = (P',X')$ iff

$$\log P + \alpha \log X > \log P' + \alpha \log X'.$$

Similarly, assume that B is priced higher than B' iff

$$\log P + \beta \log X > \log P' + \beta \log X'.$$

These relations are equivalent to the assumption that the ordering of bets according to both choice and pricing follows a multiplicative probability-value model with a power function for gains, and exponents α and β for choice and pricing, respectively. If the payoff of a bet looms larger in pricing than in choice, as implied by compatibility, β should exceed α.

To test this prediction, Tversky et al (1989) applied the above model to the data and estimated α and β separately for each subject. Note that a choice between an H-bet (P_H,X_H) and an L-bet (P_L,X_L) implies an inequality involving α. According to the above model, H is chosen

over L iff

$$\log P_H + \alpha \log X_H > \log P_L + \alpha \log X_L,$$

or equivalently whenever

$$R = \log(P_H/P_L)/\log(X_L/X_H) > \alpha.$$

Any comparison of H_i and L_i, i = 1, ..., 18, gives rise to an inequality of the form $R_i > \alpha$ or $R_i < \alpha$. For each subject, a value of α was selected so as to minimize the average squared deviations between the model and the data. Specifically, for any subject and any pair of bets (H_i, L_i) define $x_i = 1$ if $H_i >^* L_i$ and $x_i = 0$ if $L_i >^* H_i$. A value of α was selected for each subject by minimizing the quadratic loss function

$$F(\alpha) = \sum_{i=1}^{18} f(\alpha, x_i) \text{ where}$$

$$f(\alpha, x_i) = \begin{cases} x_i(\alpha - R_i)^2 & \text{if } R_i < \alpha \\ (1 - x_i)(\alpha - R_i)^2 & \text{if } R_i > \alpha. \end{cases}$$

Exactly the same procedure was used to estimate β, except that the H_i, L_i pairs were ordered by their cash equivalents, excluding ties. In accord with the compatibility hypothesis, β exceeded α for 87% of the subjects (N=179) and the difference between them was significantly positive ($p<.001$). To evaluate the adequacy of the model, the logarithm of the prices were regressed against log P and log X, separately for each subject. The median value of the multiple correlation was .95, indicating that the model provided a reasonable fit for individual data.

It should be noted that the contingent-weighting model (with $\beta > \alpha$) implies overpricing of both H and L bets. It can be shown that the predicted effect, however, is substantial for L bets and negligible for H bets. More specifically, let Y_c and Y_p, respectively, be the cash amounts that are equivalent to the bet (P,X) in choice and in pricing. It follows from the model that the discrepancy between choice and pricing, measured by $\log(Y_p/Y_c)$, is proportional to log P. It vanishes when P approaches 1, and it is large when P is small. For example, the overpricing effect implied by the model is 20 times larger when the probability of winning (P) is .1 than when it is .9. In general, P is above .9 for H bets and below .5 for L bets. The contingent-weighting model, therefore, explains the major cause of preference reversal, namely, the overpricing of L bets. Additional hypotheses are required to explain second-order effects, such as the occasional intransitivities and the slight underpricing of H bets. In the remainder of this section, we test other implications of the compatibility hypothesis in both risky and riskless choice.

Study 3: Monetary vs. nonmonetary outcomes

If preference reversals are due primarily to the compatibility of prices and payoffs, their frequency should be substantially reduced when the outcomes of the bets are not expressed in monetary terms. To test this prediction, we constructed six pairs of H and L bets, three with monetary outcomes (as in the usual PR studies) and three with nonmonetary outcomes. Two hundred and forty-eight students from the University of Oregon participated in this study. Half of the subjects first chose between all six pairs of bets and later assigned a cash equivalent to each bet. The other half of the subjects performed these tasks in the opposite order. There was no significant order effect, therefore, the data for the two groups were combined. Table 3 presents

the entire set of twelve bets and the percentage of subjects who preferred the H bet over the L bet (H > L), the percentage of subjects who assigned a higher cash equivalent to H than to L ($C_H > C_L$), and the percentage of preference reversals (PR).

Insert Table 3 about here

The data show that the percentage of choices of H over L was roughly the same in the monetary and the nonmonetary bets (63% vs. 66%), but the percentage of cases in which C_H exceeds C_L was substantially smaller in the monetary than in the nonmonetary bets (33% vs. 54%). Consequently, the overall incidence of predicted preference reversal decreased significantly from 41% to 24% ($p<.01$). Naturally, the pricing response is more compatible with monetary payoffs than with nonmonetary payoffs. Hence, the observed reduction in preference reversal with nonmonetary outcomes underscores the role of compatibility in the evaluation of options. Because even the nonmonetary payoffs can be evaluated in monetary terms, albeit with some difficulty, we do not expect the complete elimination of preference reversals in this case.

Study 4: Time Preferences

The compatibility hypothesis entails that preference reversals should not be restricted to risky choice and they should also be found in riskless options. The present study investigates this hypothesis using delayed payoffs that differ in size and length of delay (see Tversky et al, 1989). Consider a delayed payoff of the form (X, T) that offers a payment of X dollars, T years from now. Table 4 presents four pairs of options that consist of a long-term prospect L (e.g.,

Table 3

The monetary and nonmonetary bets used to test the compatibility hypothesis with the respective percentage of preferences.

		H >* L	$C_H > C_L$	PR
	Monetary Bets			
1.	H: .94 to win $3 L: .50 to win $6.50	57	26	42
2.	H: .86 to win $7.50 L: .39 to win $17	69	21	51
3.	H: .81 to win $16 L: .19 to win $56	63	51	29
	Mean	63	33	41
	Nonmonetary Bets			
4.	H: .89 to win a one-week pass good at all movie theatres in town. L: .33 to win a one-month pass good at all movie theatres in town.	65	46	30
5.	H: .92 to win an all-expenses-paid weekend at an Oregon coastal resort. L: .08 to win a one-week all-expenses-paid trip to Hawaii.	72	56	25
6.	H: .92 to win a one-week pass good at all movie theatres in town. L: .31 to win dinner for two at a very good restaurant.	62	60	16
	Mean	66	54	24

$2500, 5 years from now), and a short-term prospect S (e.g., $1600, 1 1/2 years from now).

One hundred and sixty-nine students from the University of Oregon participated in a study of choice between delayed payoffs. One-half of the subjects first chose between S and L in each pair and later priced all eight options by stating "the smallest immediate cash payment for which they would be willing to exchange the delayed payment". The other subjects performed the choice and pricing tasks in the opposite order. There were no systematic differences between the groups, so their data were combined.

Insert Table 4 about here

Table 4 presents the four pairs of options employed in this study. The table also includes, for each pair, the percentage of subjects who chose S over L ($S >^* L$), the percentage of subjects who priced S above L ($C_S > C_L$), and the percentage of PR patterns ($S >^* L$ and $C_L > C_S$). Because both the given payoffs and the stated prices are expressed in dollars, the compatibility hypothesis implies that the payoffs will be weighted more heavily in pricing than in choice. As a consequence, the preference for the short-term option (S) over the long-term option (L) should be greater in choice than in pricing. Table 4 confirms this prediction. Overall, S was chosen over L 74% of the time, but S was priced higher than L only 25% of the time, yielding 52% preference reversals, as compared with 3% reversals in the opposite direction. The application of the diagnostic analysis described earlier revealed that, as in the case of choice between simple bets, the major determinant of preference reversal was overpricing of the long-term option, as

Table 4

The options used in Study 5 and the respective percentage of preferences. The pair (X,T) denotes the option of receiving $X, T years from now.

S	L	S >* L	$C_S > C_L$	PR
(1600, 1 1/2)	(2500, 5)	57	12	49
(1600, 1/ 1/2)	(3550, 10)	72	19	56
(2500, 5)	(3550, 10)	83	29	57
(1525, 1/2)	(1900, 2 1/2)	83	40	46
Mean		74	25	52

suggested by compatibility (Tversky et al, 1989).

In the pricing task each option is evaluated singly whereas choice involves a direct comparison between options. The standard demonstrations of PR, therefore, are consistent with the alternative hypothesis that payoffs are weighted more heavily in a singular than in a comparative evaluation. To test this hypothesis against compatibility, we replicated the above study on a new group of 184 students from the University of Oregon, with one change. Instead of pricing the options, the subjects were asked to rate the attractiveness of each option on a scale from 0 (not at all attractive) to 20 (extremely attractive). If PR is controlled, in part at least, by the nature of the task (singular vs. comparative) we should expect L to be more popular in rating than in choice. On the other hand, if PR is produced by scale compatibility, there is no obvious reason why rating should differ from choice. Indeed, no discrepancy between choice and rating was observed. Overall, S was chosen over L 75% of the time (as in the original study) and the rating of S exceeded the rating of L in 76% of the cases. Only 11% of the patterns exhibited PR between choice and rating as compared to 52% between choice and pricing.

Study 3 showed that the use of nonmonetary prizes greatly reduced the amount of preference reversal whereas Study 4 demonstrated substantial preference reversal in the absence of risk. Evidently, preference reversals are controlled primarily by the compatibility between the price and the payoffs, regardless of the presence or absence of risk.

Study 5: Matching vs. Pricing

In addition to pricing and choice, options can be evaluated through a matching procedure in which a decision maker is required to fill in a missing value so as to equate a pair of options. Considerations of compatibility suggest that the attribute on which the match is made will be overweighted relative to another attribute. This hypothesis is tested in the following study, using 12 pairs of H and L bets, displayed in Table 5. In each pair, one value -- either a probability or a payoff -- was missing, and the subjects were asked to set the missing value so they would be indifferent between the two bets. Consider, for example, the bets H = (33/36; $50) and L = (18/36; $125). If we replace the 18/36 probability in L by a question mark, the subject is asked in effect "what chance to win $125 is equally attractive as a 33/36 chance to win $50?" The value set by the subject implies a preference between the original bets. If the value exceeds 1/2, we infer that the subject prefers H to L, and if the value is less than 1/2 we reach the opposite conclusion. Using all four components as missing values, we can infer the preferences from matching either the probability or the payoff of each bet. If the compatibility hypothesis applies to matching, then the attribute on which the match is made will be overweighted relative to the other attribute. As a consequence, the inferred percentage of preferences for H over L should be higher for probability matches than for payoff matches.

Two hundred subjects from the University of Oregon participated in this study. Each subject saw 12 pairs, each consisting of a high probability bet (H) and a low probability bet (L). Six of these pairs consisted of bets with relatively small payoffs; the other six pairs consisted of bets with large payoffs, constructed by multiplying the payoffs in the first six pairs by a factor of 25 (see Table 5). Each pair of bets was evaluated in four ways: direct choice, pricing of each

bet individually, matching by providing a missing payoff, and matching by providing a missing probability. Every subject performed both choice and pricing tasks, and matched either probabilities or payoffs (no subject matched both probabilities and payoffs). The order in which these tasks were performed was counterbalanced.

Insert Table 5 about here

The dependent variable of interest is the percentage of responses favoring the H bet over the L bet. These values are presented in Table 5 for all four tasks. Note that these percentages are directly observed in the choice task and inferred from the stated prices and the probability and payoff matches in the other tasks. Under procedure invariance, all these values should coincide. The overall means showed that the tendency to favor the H bet over the L bet was highest in choice (76%) and in probability matching (73%), and substantially smaller in payoff matching (47%) and in pricing (37%). These results demonstrate two types of preference reversals: i) choice versus pricing, and ii) probability matching versus payoff matching.

i) Choice versus pricing. The comparison of the results of choice and pricing in Table 5 reveals the familiar PR pattern. Subjects preferred the H bet but assigned a higher cash equivalent to the L bet. As was demonstrated earlier, this effect is due primarily to the overpricing of L bets implied by compatibility.

Table 5

Percentage of responses favoring the H bet over the L bet for four different elicitation procedures.

H		L	Choice	Probability Matching	Payoff Matching	Pricing
Small Bets:						
(35/36,$4)	or	(11/36,$16)	80	79	54	29
(29/36,$2)	or	(7/36,$9)	75	62	44	26
(34/36,$3)	or	(18/36,$6.5)	73	76	70	39
(32/36,$4)	or	(4/36,$40)	69	70	26	42
(34/36,$2.5)	or	(14/36,$8.5)	71	80	43	22
(33/36,$2)	or	(18/36,$5)	56	66	69	18
Mean			71	72	50	29
Large Bets:						
(35/36,$100)	or	(11/36,$400)	88	76	69	65
(29/36,$50)	or	(7/36,$225)	83	64	31	55
(34/36,$75)	or	(18/36,$160)	77	79	65	55
(32/36,$100)	or	(4/36,$1,000)	84	68	28	61
(34/36,$65)	or	(14/36,$210)	78	80	36	57
(33/36,$50)	or	(18/36,$125)	68	75	58	46
Mean			80	74	48	56
Overall mean			76	73	49	37

ii) Probability matching versus payoff matching. The major new result of this study concerns the discrepancy between probability matching and payoff matching. By compatibility, the dimension on which the match is made should be overweighted relative to the other dimension. Probability matching, therefore, should favor the H bet, whereas payoff matching should favor the L bet. Indeed, the tendency to favor the H bet over the L bet was much more pronounced in probability matching than in payoff matching.

Table 5 contains two other comparisons of interest: pricing versus payoff matching, and choice versus matching. Although the pricing of a bet can be viewed as a special case of payoff matching in which the matched bet has P = 1, it appears that the monetary dimension looms even larger in pricing than in payoff matching. This conclusion, however, may not be generally valid, since it holds for the small but not the large bets.

Finally, the least expected feature of Table 5 concerns the relation between choice and matching. If, relative to choice, probability matching biases the responses in favor of the H bets whereas payoff matching biases the responses in favor of the L bets, then the choice data should lie between the two matching conditions. The finding that the tendency to favor the H bet is about the same in direct choice and in probability matching suggests that an additional effect beyond scale compatibility is involved.

The missing factor, we propose, is the prominence effect demonstrated by Tversky et al (1988). In an extensive study of preference, these investigators showed that the more important attribute of an option is weighted more heavily in choice than in matching. In other words, the choice ordering is more lexicographic than that induced by matching. We have originally inter-

preted PR in terms of compatibility rather than prominence (Tversky et al., 1988), because we saw no a priori reason to hypothesize that probability is more important than money. The results of Study 5, however, forced us to reconsider the hypothesis that probability is more prominent than money, which is further supported by the finding that the rating of bets is dominated by probability (see Goldstein & Einhorn, 1987; Slovic & Lichtenstein, 1968; Tversky et. al, 1988). It appears to us now that the data of Table 5 represent the combination of two effects: a compatibility effect that is responsible for the difference between probability matching and payoff matching (including pricing), and a prominence effect that contributes to the relative attractiveness of H bets in choice. This account is illustrated in Table 6 which characterizes each of the four elicitation procedures in terms of their compatibility and prominence effects.

Insert Table 6 about here

Let us examine first the columns of Table 6, which represent the effects of the compatibility factor. Recall that the probability matching procedure enhances the significance of P and thereby favors the H bet. Analogously, the compatibility of the payoff matching and pricing procedures with the monetary outcomes enhances the significance of the payoffs and thereby favors the L bet. The choice procedure, however, is neutral with respect to the compatibility factor, hence it is expected to lie between the two matching procedures-- if compatibility alone were involved. Now consider the rows of Table 6. In terms of the prominence factor, the more important dimension (i.e., probability) is expected to loom larger in choice than in either matching procedure. Thus the tendency to choose the H bet should be greater in choice than in

Table 6

Compatibility and prominence effects for four elicitation procedures.

Prominence effect favors	*Compatibility effect favors* H	Neither	L
H		Choice	
Neither	Probability Matching		Payoff Matching, Pricing

matching, if prominence alone were involved. Table 5 suggests that both compatibility and prominence are present in the data. The finding that choice and probability matching yield similar results suggests that the two effects have roughly the same impact. It follows from this analysis that compatibility and prominence contribute jointly to the discrepancy between choice and pricing, which may help explain both the size and the robustness of the standard preference reversal. It is noteworthy that each of these effects has been established independently. The demonstrations of compatibility reported in the first part of this paper do not involve prominence, and the prominence effects demonstrated by Tversky et al (1988) do not depend on scale compatibility.

Discussion

Although the notion of compatibility has long been suggested as a possible cause of elicitation effects (see, e.g., Lichtenstein & Slovic, 1971; Slovic & MacPhillamy, 1974), this hypothesis has not heretofore been tested directly. The present investigations tested several implications of the compatibility hypothesis in studies of prediction and preference. In each of these studies, enhancing the compatibility between a stimulus attribute and the response mode led to increased weighting of that attribute. These findings indicate that compatibility plays an important role in judgment and choice. At the same time it is evident that this concept requires further theoretical analysis and empirical investigation. Implications of the present work and directions for future studies are discussed below.

The testing and application of the compatibility principle require auxiliary hypotheses about the characteristics of a stimulus attribute that make it more or less compatible with a given response mode. Many features of stimulus attributes and response scales could enhance their compatibility. These include the use of the same units (e.g., grades, ranks), the direction of relationships (e.g., whether the correlations between input and output variables are positive or negative), and the numerical correspondence (e.g., similarity) between the values of input and output variables. Although we do not have a general procedure for assessing compatibility, there are many situations in which the compatibility ordering could be assumed with a fair degree of confidence. For example, it seems evident that the prediction of market value in dollars is more compatible with a predictor expressed in dollars than with a predictor expressed in ranks. The same situation exists in the domain of perceptual-motor performance. There is no general theory for assessing the compatibility between an information display and a control panel, yet it is evident that some input-output configurations are much more compatible than others and therefore yield better performance.

Further evidence for compatibility effects in risky choice has been reported by Schkade and Johnson (1988). Using a computer-controlled experiment in which the subject can see only one component of each bet at a time, the investigators were able to measure the amount of time spent by each subject looking at probabilities and at payoffs. Their results showed that the percentage of time spent on payoffs was significantly greater in pricing than in choice. Furthermore, this pattern was particularly pronounced when the subjects produced preference reversals, and it vanished when the subjects produced consistent responses. The conclusion that subjects attend to the payoffs in pricing more than in choice supports the hypothesis that subjects focus

their attention on the stimulus components that are most compatible with the response mode. This finding is also consistent with the hypothesis that, in choice between bets, probability is perceived as more important than payoff.

In a second experiment, Schkade and Johnson (1988) compared the pricing of bets to their rating on a 100 point scale. The participants in this study expressed the ratings and the prices using an adjustable pointer. The authors observed that both the initial and the final settings of the pointer were higher for the L bet than for the H bet in pricing, and higher for the H bet than for the L bet in rating. The authors attribute the reversal of preference observed in this task to an insufficient adjustment (Slovic & Lichtenstein, 1971; Tversky & Kahneman, 1974) of the self-generated anchors. The productions of these anchors, however, appears to be governed by compatibility. Note that the response scale in the pricing task ranges from 0 to the positive payoff, whereas the range of the rating scale (0 to 100) matches the probability scale. By compatibility, the payoff is expected to loom larger in pricing than in rating, and the probability is expected to loom larger in rating than in pricing. The notion that the bounded rating scale is more compatible with probability than with money, supported by the process data of Schkade and Johnson, may explain the finding (Goldstein & Einhorn, 1987) that the preference for the H bet over the L bet is stronger in rating than in choice, despite the procedural similarity between rating and pricing. An alternative explanation of this result that attributes PR to the mapping of subjective value onto the response scale rather than to the compatibility between stimulus components and response modes was proposed by Goldstein and Einhorn (1987). Their model can accommodate reversals of preferences, but it does not predict the variety of compatibility effects described in the present paper.

Recent results reported by Delquie´ and de Neufville (1988) are also consistent with the compatibility hypothesis. These authors employed a double-matching procedure, devised by Hershey and Schoemaker (1985), in which subjects first determine the missing value (e.g., the probability of winning) of an option that would make it equivalent to a second option. Later, the subjects are presented with the option they constructed and they now have to determine the missing value (e.g., the payoff) of the second option that would make the two options equally attractive. If procedure invariance holds, the latter match should coincide with the given value of the second option. Using both risky and riskless options, Delquie´ and de Neufville found systematic violations of procedure invariance, which imply that the matched attribute is weighted more heavily than the other attribute -- as predicted by compatibility. These findings confirm, in a double-matching design, the conclusion of Experiment 5 and of Tversky et. al (1988) that were based on a choice-matching design.

The compatibility notion discussed in this paper concerns the correspondence between the scales in which the inputs and outputs are expressed. In a previous paper (Tversky et al., 1988), we have explored a more abstract notion of compatibility that was later called "strategy compatibility" by Fischer and Hawkins (1988). To introduce this concept, we distinguished between qualitative and quantitative choice strategies. Qualitative strategies (e.g., dominance and minimax) are based on purely ordinal criteria whereas quantitative strategies (e.g., multiattribute utility theory) are based on trade-offs or weighting of the dimensions. We proposed that the qualitative strategy of selecting the option that is superior on the more important dimension is more likely to be employed in the qualitative method of choice, whereas a quantitative strategy based on the trade-offs between the dimensions is more likely to be used in the

quantitative method of matching. In this sense, the prominence effect may be attributable to the compatibility between the nature of the task and the nature of the strategy it invokes. For further discussion of strategy compatibility and its relationship to scale compatibility, see Fischer and Hawkins (1988).

Compatibility, like anchoring, can have a powerful effect on prediction and preference, yet people appear to have little or no conscious awareness of it, either inside or outside the laboratory. Such bias seems to operate at a very elementary level of information processing and it is doubtful whether it can be eliminated by careful instructions, or by monetary payoffs. Indeed, the use of incentives to promote careful responses has had little influence on the prevalence of preference reversals (Slovic & Lichtenstein, 1983).

The effects of compatibility described in this chapter represent a major source of violations of procedure invariance, namely the requirement that normatively equivalent elicitation procedures should yield the same ordering of options or events. The failure of procedure invariance complicates the task of the practitioner and the theorist alike. From a practical perspective, the present findings underscore the lability of judgments and choices, and make the elicitation task quite problematic. If the decision maker's response depends critically on the method of elicitation, which method should be used, and how can it be justified? At the very least, we need to use multiple procedures (e.g., choice, pricing, rating) and compare their results. If they are consistent, we may have some basis for trusting the judgment, if they are not, further analysis is required.

The assumption of procedure invariance plays an essential role in theories of rational choice. Behavioral research has also demonstrated consistent violations of description invariance by showing that different descriptions of the same decision problem can give rise to systematically different choices. Thus, alternative framings of the same options (e.g., in terms of gains vs. losses, or in terms of survival rates vs. mortality rates) produce predictable reversals of preference (Tversky & Kahneman, 1986). These failures of description invariance, induced by framing effects, and the failures of procedure invariance, induced by elicitation effects, represent deep and sweeping violations of classical rationality.

Attempts to describe and explain these failures of invariance require choice models of much greater complexity. To account for violations of description invariance, it seems necessary to introduce a framing process, including the determination of a reference point, which takes place prior to the valuation of prospects (Kahneman & Tversky, 1979). To account for violations of procedure invariance, it seems necessary to introduce multiple preference orders (obtained from choice, matching or pricing) and a contingent weighting model (Tversky et al., 1988) in which the tradeoff among attributes is contingent on the method of elicitation. These developments highlight the discrepancy between the normative and the descriptive approaches to decision making. Because invariance--unlike independence or even transitivity--is normatively unassailable and descriptively incorrect, it may not be possible to construct a theory of choice that is both normatively acceptable and descriptively adequate.

References

Brainard, R. W., Irby, T. S., Fitts, P. M., & Alluisi, E. (1962). Some variables influencing the rate of gain of information. *Journal of Experimental Psychology, 63,* 105-110.

Delquie', P., & de Neufville, R. (1988). *Response-modes and inconsistencies in preference assessments.* Unpublished manuscript, Massachusetts Institute of Technology.

Fischer, G. W., & Hawkins, S. A. (1988). *Preference reversals in multiattribute decision making: Scale compatibility, strategy compatibility, and the prominence effect.* Unpublished manuscript, Carnegie-Mellon University.

Fitts, P. M., & Seeger, C. M. (1953). S-R compatibility: Spatial characteristics of stimulus and response codes. *Journal of Experimental Psychology, 46,* 199-210.

Goldstein, W. M., & Einhorn, H. J. (1987). Expression theory and the preference reversal phenomena. *Psychological Review, 94,* 236-254.

Grether, D. M., & Plott, C. R. (1979). Economic theory of choice and the preference reversal phenomenon. *American Economic Review, 69,* 623-638.

Hershey, J., & Schoemaker, P. (1985). Probability versus certainty equivalence methods in utility measurement: Are they equivalent? *Management Science, 31,* 1213-1231.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica, 47,* 263-291.

Lichtenstein, S., & Slovic, P. (1971). Reversals of preference between bids and choices in gam-

bling decisions. *Journal of Experimental Psychology, 89*, 46-55.

Lichtenstein, S., & Slovic, P. (1973). Response-induced reversals of preference in gambling: An extended replication in Las Vegas. *Journal of Experimental Psychology, 101*, 16-20.

Schkade, D. A., & Johnson, E. J. (1988). *Cognitive processes in preference reversals.* Working paper, Department of Management, University of Texas.

Slovic, P., Fischhoff, B., & Lichtenstein, S. (1982). Response mode, framing, and information-processing effects in risk assessment. In R. Hogarth (Ed.), *New directions for methodology of social and behavioral science: Question framing and response consistency*, no. 11 (pp. 21-36). San Francisco, CA: Jossey-Bass.

Slovic, P., & Lichtenstein, S. (1968). The relative importance of probabilities and payoffs in risk-taking. *Journal of Experimental Psychology Monograph Supplement, 78*(3), part 2 (b).

Slovic, P., & Lichtenstein, S. (1971). Comparison of Bayesian and regression approaches to the study of information processing in judgment. *Organizational Behavioral and Human Performance, 6*, 649-744.

Slovic, P., & Lichtenstein, S. (1983). Preference reversals: A broader perspective. *American Economic Review, 73*, 596-605.

Slovic, P., & MacPhillamy, D. (1974). Dimensional commensurability and cue utilization in comparative judgment. *Organizational Behavior and Human Performance, 11*, 172-194.

Tversky, A. (1977). Features of similarity. *Psychological Review, 84*, 327-352.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Sci-*

ence, 185, 1124-1131.

Tversky, A., & Kahneman, D. (1986). Rational choice and the framing of decisions. *The Journal of Business, 59* (4), Pt. 2, 251-278.

Tversky, A., Sattath, S., & Slovic, P. (1988). Contingent weighting in judgment and choice. *Psychological Review, 95* (3), 371-384.

Tversky, A., Slovic, P., & Kahneman, D. (1989). *The causes of preference reversal.* Unpublished manuscript, Stanford University.

Wickens, C. D. (1984). *Engineering psychology and human performance.* Columbus: Merrill.